Beyond gene expression

Beyond gene expression

Novel methods and applications of transcript expression analyses in RNA-Seq

Cover design by Iftikhar Hussain (Iftikhar159; www.fiverr.com)
Printed by Gildeprint

Beyond gene expression

Novel methods and applications of transcript expression analyses in RNA-Seq

Dissertation
to obtain the degree of Doctor at Maastricht University,
on the authority of the Rector Magnificus Prof. dr. Rianne M. Letschert
in accordance with the decision of the Board of Deans,
to be defended in public on

Wednesday 3rd of March 2021 at 16.00 hours (CEST)

by

Rajinder Gupta

born on April 4, 1989
in Jammu, J&K, India

Supervisor

Prof. dr. Jos C.S. Kleinjans

Co-supervisor

Dr. Florian Caiment

Assessment Committee

Prof. dr. Andre Dekker (chair)

Prof. dr. Chris Evelo

Prof. dr. ir. Ralf Peeters

Prof. dr. Thomas Hartung, Johns Hopkins Bloomberg School of Public Health

Dr. Ralf Herwig, Max Planck Institute for Molecular Genetics

The research described in this thesis was conducted at GROW School of Oncology and Developemntal Biology of Maastricht University. This work was funded by EU-ToxRisk project (An Integrated European "Flagship" Program Driving Mechanism-Based Toxicity Testing and Risk Assessment for the 21st Century) funded by the European Commission under the Horizon 2020 program (Grant Agreement No. 681002).

Table of contents

Chapter 1

General Introduction

Introduction

Background

With the invention of Sanger sequencing (also known as the first generation of sequencing) in 1977, the landscape of the DNA sequencing changed forever [1-3]. The Sanger sequencing works by selective incorporation of chain-terminating dideoxynucleotides using DNA polymerase during *in vitro* DNA replication. For nearly three decades, Sanger sequencing was the go to method for sequencing [4]. One of the biggest application of Sanger sequencing was in the Human Genome Project (HGP). For the very first time, the code that makes us human was not a secret anymore. Though the HGP was a great success, the time (~13 years), money ($2.7 billion), and the manpower (numerous researchers and technicians from 20 institutions across the globe) it required was overwhelming [https://www.genome.gov/human-genome-project/Completion-FAQ].

The challenges in Sanger sequencing paved the way for next-generation sequencing (NGS) and they came into existence with the onset of the twenty-first century. NGS addressed the limitations of Sanger sequencing such as time and cost, sequencing only short pieces of DNA (300-1000 bases), the poor quality of first and last 15 to 40 bases [5], and the quality degrades for longer sequences. The NGS technologies could sequence large fragments, made comparatively fewer errors, were faster and cheaper, and could sequence multiple samples in one go. The size of data generated by a single run of NGS can range from a few Gbs to thousands of Gbs (NovaSeq 6000; output range of 4800 - 6000 Gb) and all of this can be achieved in few hours to few days [https://www.illumina.com/systems/sequencing-platforms/novaseq.html]. The projection for genomics data storage stands at 2–40 EB/year (1 EB = 10^{18} bytes or 1 billion gigabytes) by 2025 [6].

Besides all the knowledge which was achieved with the commencement of the HGP, one important piece of theory, “one gene-one enzyme”, as proposed by George Beadle and Edward Tatum in 1941 [7] did not hold true anymore. The number of genes in humans was established to be ~20000 while the number of proteins was found to be much higher. Uniprot, one of the most popular protein database, houses 20,368 manually annotated and 171,915 computationally annotated human proteins (assessed on 29/07/2020) [8]. Similarly, Ensembl houses 249,740 transcripts for human (GRCh38.p13) (assessed on 29/07/2020) [9]. The credit for the existence of so many proteins or transcripts from so fewer genes can

be accredited to alternative splicing, single amino acid polymorphisms (SAPs) arising from nonsynonymous single-nucleotide polymorphisms (nsSNPs), and posttranslational modifications (PTMs) [10, 11]; as many as 100 different proteins can potentially be produced from a single gene.

The transcriptome comprises of all types of RNA transcripts. The transcriptomic landscape goes under tremendous changes owing to different growth phases, diseases, lifestyle, environmental factors, etc. The identification and quantification of the various types of transcripts can provide specific signatures for diseases and treatment. A detailed classification of the transcripts, adapted from Ensembl transcript classification (https://vega.archive.ensembl.org/info/about/gene_and_transcript_types.html), is presented below.

1. **Protein coding**: The transcripts that are translated into a functional protein and are further classified as follows:

a. Known protein coding: The transcripts that are 100% identical to a RefSeq NP (curated non-redundant sequence database of genomes, transcripts, and proteins) [12] or Swiss-Prot entry (manually annotated and reviewed Uniprot entry).

b. Novel protein coding: The transcripts that share more than 60% length with known coding sequence from a RefSeq or Swiss-Prot or have cross-species/family support or domain evidence.

c. Putative protein coding: The transcripts that share less than 60% length with known coding sequence from a RefSeq or Swiss-Prot, or have an alternative first or last coding exon.

d. Nonsense mediated decay (NMD): The coding sequence of a transcript carrying a premature termination codon (PTC) that finishes more than 50 base pairs (bp) from a downstream splice site. The transcripts with PTC are a result of nonsense or frameshift mutations in endogenous genes, pseudogenes, or intron retention or inclusion of PTC-containing exons from alternative splicing [13-15]. NMD transcripts are recognized and subsequently degraded to avoid deleterious effects for the organism [16].

e. Nonstop decay (NSD): Transcripts that have polyadenylation (polyA) features without an in-frame stop codon, i.e. a non-genomic polyA tail attached directly to the CDS without 3' untranslated region (UTR). NSD transcripts are formed when

transcription aborts abruptly, polyadenylation occurs prematurely, or through point mutations that disrupt the stop codon [17]. These transcripts are subject to degradation.

2. **Processed transcripts**: The transcripts that do not contain an open reading frame (ORF) and are further divided into three major categories.

a. Long non-coding RNA (lncRNA): The transcripts that do not code for proteins and have a length of more than 200 nucleotides. They are sub-classified into one of the following types:

i. 3-prime overlapping ncRNA: The transcripts that have di-tag and/or published experimental data which strongly support the existence of long non-coding transcripts that overlaps the 3' UTR of a protein-coding locus on the same strand.

ii. Antisense: The transcripts that overlap the genomic span (i.e. exon or introns) of a protein-coding locus on the opposite strand. They are complementary to a protein coding messenger RNA (mRNA) and are known to play an important role in the regulation of gene expression [18].

iii. lincRNA (long interspersed ncRNA): The long intervening/intergenic non-coding RNAs that do not overlap protein-coding genes. They lack coding potential and may not be conserved between species. Some known functions of lincRNA are remodeling chromatin and genome architecture, RNA stabilization, and transcription regulation, including enhancer-associated activity [19].

iv. Retained intron: An alternatively spliced transcript that has an intronic sequence relative to other coding variants.

v. Sense intronic: A long non-coding transcript in introns of a coding gene that does not overlap any exons.

vi. Sense overlapping: A long non-coding transcript that contains a coding gene in its intron on the same strand.

vii. Macro lncRNA: Un-spliced lncRNAs that are several kb in size. They are particularly found in imprinted gene clusters and have been shown to silence various cis-genes in imprinted gene clusters [20].

b. ncRNA (non-coding RNA): A RNA molecule that is not translated into a protein and is of size less than 200 nucleotides. The ncRNA are further classified as below:

i. miRNA (microRNA): These are small ncRNA with a size of ~22 nucleotides. Their role in gene silencing and translational repression by binding to target mRNAs has presented them as potential therapeutic targets [21]. Lately, their significance in biological processes, development, and progression of various diseases in thoroughly investigated [22, 23].

ii. piRNA (piwi-interacting RNA): The small ncRNA that binds to the piwi-subfamily *Argonaute* proteins and has a size of 26-31 nucleotides. They are crucial for genome defense, and it is achieved by transcriptional and posttranscriptional silencing of the transposable elements [24].

iii. siRNA (small interfering RNA): They are double-stranded non-coding RNA molecules, typically 20-27 base pairs in length. They are also known as short interfering RNA or silencing RNA. They are a powerful tool for the regulation of gene expression and has been utilized as a therapeutic agent against various diseases [25]. It binds to the complementary mRNA and induces mRNA cleavage, hence preventing translation [26].

iv. snRNA (small nuclear RNA): In eukaryotic cells, small RNA molecules found within the splicing speckles and Cajal bodies of the cell nucleus are snRNA. They have an average length of ~150 nucleotides. They are always associated with a set of specific proteins, and the complexes hence formed are referred to as small nuclear ribonucleoproteins (snRNP, pronounced as "snurps"). They are subunits of the spliceosome, involved in catalyzing the pre-mRNA splicing [27].

v. snoRNA (small nucleolar RNA): The small ncRNA that guides the chemical modifications (methylation and pseudouridylation) of other RNAs, such as ribosomal RNAs, transfer RNAs, and small nuclear RNAs [28, 29].

vi. tRNA (transfer RNA): An adaptor molecule composed of RNA, having a length of 76-90 nucleotides. It acts as the physical link between the mRNA and the amino acid sequence of proteins.

vii. vaultRNA: The short non-coding RNAs, ~100 nucleotides long, that form part of the vault ribonucleoprotein complex. They are shown to be involved in nucleocytoplasmic transport [30], riboregulation [31], and drug-resistance [32].

c. rRNA (ribsosomal RNA): rRNA is a ribozyme that is the primary component of ribosomes and carries out protein synthesis. The ribosomes are made of two subunits: the large ribosomal subunit (60S) and the small ribosomal subunit (40S).

The large subunit consists of three rRNA molecules (5S: 121 nucleotides, 5.8S: 156 nucleotides, 28S : 5070 nucleotides) and the small subunit is composed of a single rRNA molecule (40S) of size 1869 nucleotides [33]. The ribosome binds to both mRNA and tRNA to facilitate the translation of codons in the mRNA into amino acids. A polysome is formed when a single mRNA molecule is translated simultaneously by multiple ribosomes, creating multiple copies of the protein [34].

3. **Pseudogene**: A paralogous gene that has one or more of the further listed characteristics: missing promoter, start codon, or introns, frameshift, premature stop codon, and partial deletion. Sometimes these entries have an intact coding sequence or a truncated ORF, in which case there is other evidence used (for example genomic polyA stretches at the 3' end) to classify them as a pseudogene. These can be further classified as follows:

a. Processed pseudogene: Pseudogene that appears to have been produced by the integration of a reverse transcribed mRNA into the genome.

b. Unprocessed pseudogene: Pseudogene that shows evidence of loss of function, but has a exon-intron structure.

c. Polymorphic pseudogene. Pseudogene owing to a single nucleotide polymorphism, a deletion, or an insertion polymorphism but in other individuals / haplotypes / strains the gene is translated.

d. Unitary pseudogene: A species-specific unprocessed pseudogene without a parent gene, as it has an active orthologue in another species.

e. IG pseudogene: An inactivated immunoglobulin gene.

Though the presence of various types of transcripts is well established, the research community is primarily focused on studying gene expression. The gene-based analyses are an ideal choice to study gene regulation, gene expression, genetic networks, gene fusions, and genetic mutations. However, while assessing the functional capacity of the biological systems, the gene-based analysis has limited applicability due to its capacity to code for several protein coding and non-coding transcripts. The gene expression is an ensemble of expression of various transcripts that originate from the given gene (Figure 1). If the expression of one transcript is decreased and another increased by the same magnitude, the global gene expression might seem unperturbed. A gene-expression based study masks the changes occurring at the transcript expression level.

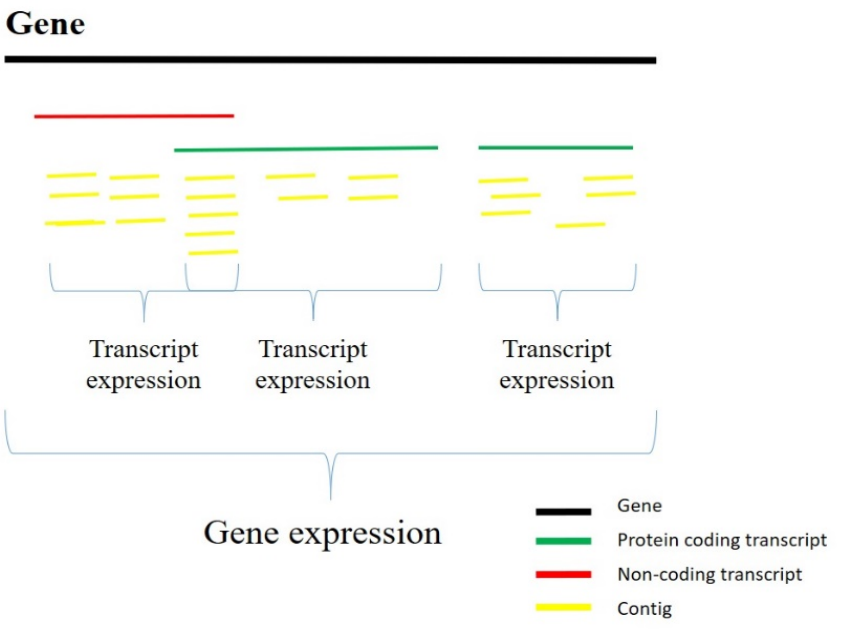


***Figure 1**: Illustration of gene and transcript expression. The contigs obtained from sequencing are mapped to the regions corresponding to different transcripts. The summation of these contigs give the transcript expression, the contigs that are shared between different transcripts are resolved using various techniques. These transcripts can be protein coding or non-coding. From the transcript expression, then the gene expression is obtained by the accumulation of expression of all transcripts for that gene.*

The vast repertoire of transcripts found in almost all living systems can be attributed to various transcription start sites (TSS) [35] and alternative splicing (AS) [36]. The presence of multiple transcription start sites explain the pervasive transcription and may lead to translational differences [35, 37]. Consecutively, a large chunk of transcript repertoire is attributed to AS. Through AS, from one pre-mRNA multiple spliced mature mRNA can be produced. AS functions by inserting and deleting important functional domains that are encoded by alternatively spliced exons [36].

Omnipresence of gene-based analyses

There are multiple reasons why gene expression based analyses in transcriptomics studies are preferred, the main arguments being non-exhaustive transcript

identification, the unknown function of the transcripts, and the prevalence of gene based biological databases and tools.

The knowledge of the existence of different transcripts is known for long; however, there is still a lot to be learned about their structure, functions, and origin. Over the last five decades, we have not yet managed to identify and enumerate all transcripts for any species, let alone humans. However, with the introduction of NGS technologies and advancements in bioinformatics, the discovery of the transcripts has gained pace in the last two decades [38, 39]. Large-scale genomics and transcriptomics studies have added immense knowledge to our understanding of transcripts in recent years. Nearly 250k transcripts in humans have been identified so far [9], yet there are more that are still to be found.

For many protein coding transcripts, the functions and/or ontologies are established, however in the case of non-coding transcripts limited or no knowledge of their function is available. From transcriptomics data analyses highly perturbed transcripts can be identified, however, if they cannot be connected to phenotypes, diseases, or molecular dysregulations, conclusions cannot be derived on functional changes. The biological databases, most of them, if not all, focus on genes and proteins. Due to the over-simplistic notion of "*one gene one protein one function*", it is more convenient to show the pathways (KEGG [40], Reactome [41] or, BioCyc [42]), ontologies (Gene ontology resource [43]), functions (Entrez-gene [44], Uniprot [8]), and AOPs (adverse outcome pathways OECD knowledgebase [45]) in terms of proteins or genes. Similar is the case with tools that are used to analyze transcriptomics data.

However, through an application programming interface (APIs) and additional on-web functionality some tools, web applications or databases allow performing transcript-based analysis such as Biomart [46], Uniprot. Nevertheless, in all the cases, the loss of data is inevitable and the conflicting cases of 1:many and 1:none identifier mappings across databases cannot be resolved. Many tools developed for gene expression analyses can be used for transcript expression by first converting the identifiers (for instance, using Biomart) but it requires some basic programming skills.

Besides APIs and identifier conversions, work has been done to analyze the transcript expression data and present the results in the form of the genes by selecting the main transcript of the gene. Two approaches that are usually followed

to achieve this are: selecting the longest protein coding transcript or APPRIS defined principal isoform. APPRIS (short for annotating principal splice isoforms) takes into account protein structural and functional features and information from cross-species conservation to select the principal isoform for the gene [47]. Selecting one transcript from a gene results in overlooking other transcripts from the given gene; however, close to 95% of human multi-exon genes undergo alternative splicing [48] and may hold important biological signals. Thus, emphasizing the importance of analyzing all transcripts rather than selecting one representative of the gene.

With the increase in our knowledge about transcripts, novel methods and tools to analyze, and knowledgebase of the transcriptomics data have been developed. GRO-seq is a method developed to measure nascent RNA and to study the function and mechanism of action of non-coding RNAs [49]. TumorFusions is a database of cancer-associated transcript fusions [50]. A novel approach to calculate the capacity of a transcript that it exerts in a cell as an enzyme or a protein function after being translated [51]. However, multiple new applications have been developed around the transcripts and their expression; there is a fundamental problem in transcript level expression data analyses.

The limitation with transcript-based analyses

Certain challenges with the transcript-based analyses have already been discussed in the previous section however, there is also a fundamental problem associated with analyzing transcript expression data. As already mentioned, one protein coding transcript will code for a single type of protein, barring the posttranslational modifications. However, different transcripts can code for similar proteins, owing to codon degeneracy. While such transcripts have a similar coding sequence (CDS), they differ from each other in terms of 3’ or 5’ UTR, introns, and transcription start sites [52, 53]. On certain occasions, they may be formed due to premature transcription termination [54].

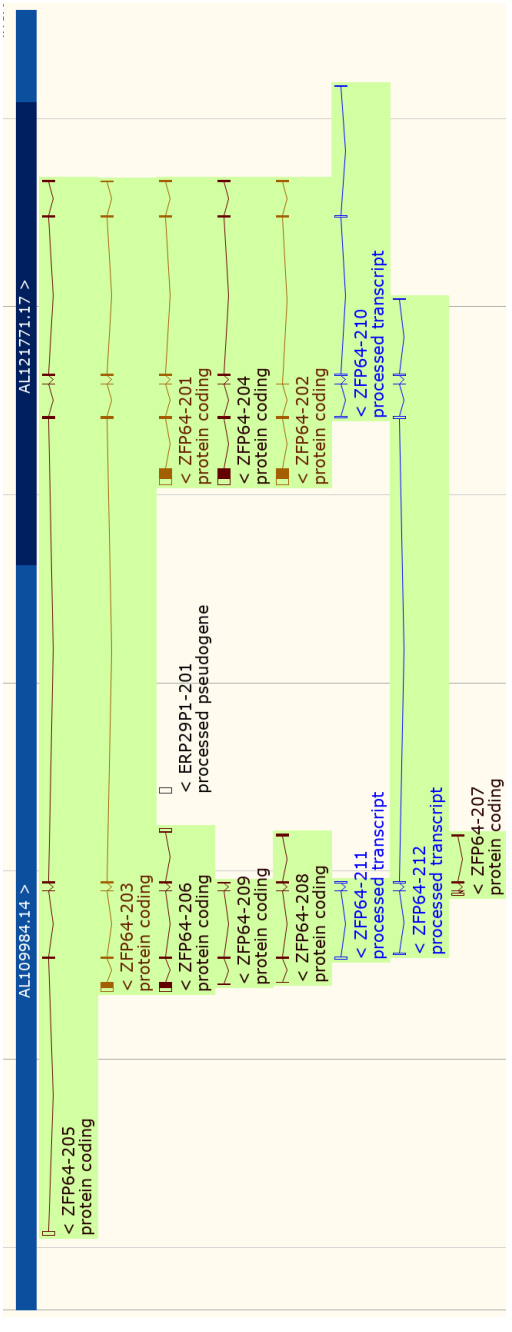


***Figure 2**: Various transcripts from ZFP64 gene are shown. The transcripts that originate from overlapping regions of the gene are mapped to the same Uniprot identifier. The transcripts ZFP64-201, ZFP64-202, ZFP64-203, ZFP64-204, and ZFP64-206, all overlap certain regions of ZFP64-205, however, the length of the transcripts is highly variable. All these transcripts are annotated to a single Uniprot identifier – Q9NTW7.*

A careful look at the data from Ensembl highlights that different protein coding transcripts from a gene are mapped to the same Uniprot identifier and have high sequence similarity. With respect to the position on the chromosomes, some

transcripts overlap and others did not. In the case of the gene ZFP64 (Zinc finger protein 64) (Figure 2), six protein coding transcripts are mapped to the same Uniprot Identifier (Q9NTW7). ZFP64-205 spans over the longest chromosomal stretch among the six transcripts, from 51693496 to 51705190. The other five transcripts: ZFP64-201, ZFP64-202, ZFP64-203, ZFP64-204, and ZFP64-206, all overlap certain regions of the ZFP63 transcript. The ZFP64-206 transcript is mapped to a different region on the chromosome than ZFP64-201, ZFP64-202, and ZFP64-204, however, they are all mapped to the same Uniprot identifier implying that they have a similar function(s). Similar trends can be observed across different genes – transcripts from different genes mapping to the same Uniprot identifier. Through this observation, it can be implied that the transcripts coding for the same/similar proteins may originate from the same or different genes. A preliminary analysis of the Ensembl and Uniprot identifier mapping revealed that in ~40k instances, two or more transcripts originating from the same or different Ensembl genes mapped to the same Uniprot identifier.

Multiple different transcripts are capable enough to code for the same proteins and hence they should be analyzed together as one entity rather than considering all transcripts individually. To achieve this, identification of transcripts that may code for the same/similar proteins has to be done thoroughly. Studying the protein structure can help in identifying proteins that have similar functions. Through the structure of the proteins, structural and functional domains can be identified, as well as various binding sites (major and minor grooves) can be located. Using this information, the similarity of the protein functions can be predicted. However, due to the scarcity of high-quality protein structure data – PDB lists a total of 48783 human protein structures, of which only 17.9K with a resolution below 2 Å are present [55] and the inefficiency of the *in silico* approaches to predict the 3D structure of proteins correctly [56], limits our ability to compare the protein 3D structures. Analyzing amino acid sequences and secondary structures of the proteins can be a practical option to provide a function-based comparison.

The omics, protein sequence, and structural data have accumulated into gargantuan amounts of data; resulting in the handling and analysis of the data overwhelming. However, the advances in the field of machine learning over the past decade have provided great promise.

Machine learning and its applications

Any set of techniques that enable computers to mimic human behavior are termed as artificial intelligence (AI) and the specific branch of AI that uses data to train the computers to learn and identify the patterns, and infer decisions without (or with least) human intervention are termed as machine learning (ML). Interest in machine learning has gained attention due to the growing volumes and varieties of data, cheaper and more powerful computational processing, and affordable data storage. Now it has become possible to create models quickly and automatically which can analyze bigger and complex data and produce faster and accurate results.

Primarily machine learning approaches can be divided into two categories: supervised and unsupervised. In supervised learning, the patterns are learned from the "labeled" data and in case of unsupervised learning from the "unlabeled" data. The labeled data has defined categories, for instance, healthy and diseased. To generate the models using the supervised algorithms, the knowledge of the categories of the data is used. Supervised learning algorithms are further split into classification and regression; and unsupervised learning into clustering and association, based on the type of tasks they perform. In classification, the given data is categorized into predefined groups, e.g. if a drug is toxic or non-toxic. The classification can be binary or multiclass depending on the number of categories into which the algorithm divides the data. Regression finds application in predicting the future value of a variable based on the knowledge that has been acquired from data, for instance, given the expression of a gene for time points t_n, t_{n+1}, and t_{n+2}; the expression of the gene at t_{n+3} can be predicted. Clustering is similar to classification as it also categorizes the data into different categories except that the categories are not predefined (unlabeled data). One of the most popular examples of clustering in omics data studies is of principal component analysis (PCA) plots. PCA tries to find a dimension that can separate the data into various clusters. Lastly, the association is used to look for rules to establish relationships amongst various variables in unlabeled data, for example, building network graphs from gene or transcript expression data. The association helps in discovering new relationships between the variables.

ML algorithms are capable of handling huge amounts of data, limited only by the computational power and time. Owing to ML's ability to handle huge data, learn and identify hidden patterns – quickly and automatically; supervised and

unsupervised algorithms have been applied to a vast array of biomedical data with great success. One of the major application has been improvements in the underlying processes of drug discovery such as target identification and validation [57, 58], and small-molecule design and optimization [58, 59]. In a recent initiative, the sensitivity of the drugs was also predicted using ML on a cohort of genomics, epigenomics, and proteomics profiling data sets measured in human breast cancer cell lines [60]. Moreover, our understanding of the transcriptome landscape is attributed to ML's ability to accurately predict alternative splicing signals [61]. Image analyses for abnormal tissue detection, patient stratification, and disease diagnosis or prediction has also been done with great accuracy using ML [58]. Another major aspect of disease diagnosis where recent novel findings are all attributed to the ML algorithm is the identification of biomarkers [62-65].

Biomarkers

Biomarkers (short for biological marker) are defined as distinctive and measurable biological characteristics that can be used to evaluate normal or pathogenic biological processes [66], A biomarker must be accurately measurable by some technology and should be reproducible [67]. Many molecules can be used as biomarkers, such as nucleic acids (DNA and RNA), proteins, peptides, lipids, antibodies, several metabolites, and other small molecules [66, 68, 69]. These biomolecules can be detected and quantified using various techniques from the tissues and/or bio-fluids. The biomarkers can be classified as predisposition, diagnostic, prognostic, and predictive, based on the clinical information they provide [69-72] (Figure 3). Though discovery of novel and potent biomarkers has gained momentum with the evolution of omics, the discovery of the biomarkers is not limited to NGS data.

For instance, various assays were performed to discover the predisposition biomarkers for formaldehyde exposure. Formaldehyde is a carcinogen and its exposure is both environmental and occupational. An increased risk for cancer development among workers exposed to formaldehyde has been revealed by several epidemiological studies [73-75]. It has been shown that there were alterations in the percentage of T-cytotoxic lymphocytes, NK cells, and B-lymphocytes between control and test groups. Moreover, polymorphisms in CYP2E1, GSTP1, and FANCA genes were demonstrated to be associated with an

increased genetic damage [75]. Similarly, the search for predisposition biomarkers in case of developing leukopenia (white blood-cell deficiency) was studied in northeast Brazil using Maximum likelihood on data from several assays. It was extrapolated that GSTT1 and/or GSTM1 can be used as susceptibility biomarkers for leukopenia [76].

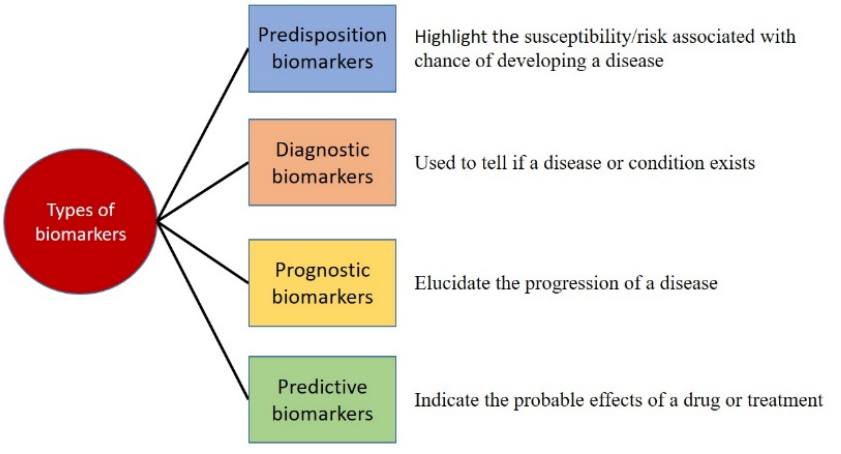


***Figure 3**: Different types of biomarkers. Four types of biomarkers, based on the type of clinical information they provide, namely predisposition, diagnostic, prognostic, and predictive biomarkers.*

Many less studied genes/proteins, for which the functions might not be known, can be projected as potent biomarkers when transcriptomics data is used. As in the case of colorectal cancer, the integration of multi-platform transcriptomics data together with several ML algorithms allowed the discovery and validation of diagnostic biomarkers. While poor disease-free survival was shown with the overexpression of TGFBI and S100A2, the down-regulation of NR5A2, SLC4A4, and CD177 was linked to the worse overall survival of the patients [77]. In another study, several ML algorithms and different feature selection techniques were used to predict biomarkers from multi-platform transcriptomics data. Three genes, namely, FCN3, CLEC1B, and PRC1 were identified as hepatocellular carcinoma prognostic indicators for various types of survival of HCC patients [78]. There are also some examples where both, the transcriptomics data and various assays were used. One such example is the discovery of novel diagnostic markers, prognostic markers, and therapeutic targets from soft tissue sarcomas [79].

Numerous studies have been conducted in search of novel biomarkers for various diseases and conditions using ML algorithms [80-83]. A common feature among all these studies is their focus on finding the "gene" biomarkers. However, as discussed before, gene expression can be misleading at times, and hence the homogeneous transcript expression should be a preferred choice for biomarker discovery. Some recent work has highlighted the added advantage of using transcripts as biomarkers with respect to increased prediction accuracy. In a study to identify novel biomarkers for increased risk to develop metabolic disorders, a set of transcript-based biomarkers indicative of a predisposition to metabolic syndrome-related alterations were identified in rats models. Among different biomarkers found in the study, NPC1 was further validated in humans due to its involvement in both lipid and glucose metabolism, as well as insulin sensitivity. Decreased NPC1 transcript levels in peripheral blood cells were observed and it was projected as a candidate biomarker of increased risk for impaired metabolic health in humans [84]. In another study, in an attempt to find better biomarkers to measure the effects downstream of FGFR pathway inhibition, four transcripts were identified from the genes: DUSP6, ETV5, YPEL2, and EGR1 [85]. It was demonstrated that these transcript biomarkers were more robustly modulated by FGFR inhibition than some conventional downstream signaling protein biomarkers.

The identification of biomarkers aid in timely prognosis of deviant behavior of the system. However, to investigate probable reasons for the development of the diseases, the knowledge of the cohort of binding partners of the perturbed protein is pertinent. The binding partners can potentially form protein complexes that may be involved in several functions. Once the binding partners and functions are identified, the information might allow looking for techniques and approaches to manipulate the binding by pharmacological inhibitors to help cure the disease [86].

Protein complexes – the molecular machinery

The intracellular environment is crowded and all different types of biomolecules come in close physical contact with each other [87, 88]. The frequency, specificity, affinity, and duration of these interactions are highly variable [86]; based on these factors the complexes can be divided into stable (or permanent) and transient complexes (Figure 4). The stable complexes demonstrate assemblies that have a long half-life, bigger interaction interface, and the interactions are found in the

molar range (M range) [89, 90]. Such complexes have been identified by experimental characterization and have well-defined molecular functions. On the other hand, transient complexes are short-lived and are usually weak (µM range). They arise as a result of posttranslational modification of one or both proteins involved in the interaction [89, 90]. The transient complexes have important functions in cellular signaling [91, 92]. In a transient protein-protein interaction, one or both proteins in the interaction undergo conformational changes, often to reveal a binding site for the next interacting protein [86]. However, there are some of the other transient complexes that have little biological relevance and are formed mostly due to intracellular crowding [93, 94].

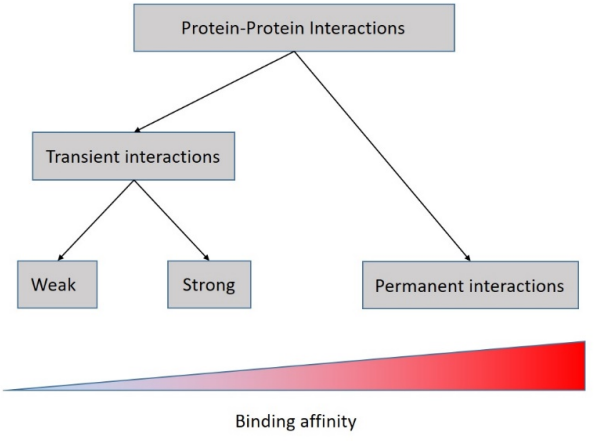


***Figure 4**: Different types of protein-protein interactions based on their binding affinities.*

The protein complexes can also be divided based on the types of subunits they consist of: homomeric or heteromeric [95]. As the name suggests, homomeric complexes are made up of the same subunits and heteromeric complexes from different subunits. Most known protein complex structures are homomeric [96, 97], however, this observation could be strongly biased due to the inclination towards studying individual proteins [98].

Currently, there are four main experimental methods for the detection of protein complexes: X-ray crystallography, nuclear magnetic resonance (NMR), various mass-spectrometry techniques such as native, cross-linked (CX or XL), and ion

mobility, and protein microarray. Besides these experimental methods, there are also various computational methods for protein complex prediction (Table 1). All computational methods take advantage of the information on the structure and topology of the given protein-protein interaction network (PPIN) [99]. Some of the computational methods make use of the unweighted PPIN whereas others employ weighted PPINs to model dense subnetworks as complexes. The discussion on the application, advantages, and limitations of experimental and *in silico* approaches is beyond the scope of this thesis and can be reviewed elsewhere [99].

***Table 1**: Various computational methods available for protein complex prediction. Adapted from Zahiri J, et al, 2019 [99].*

Type	**Sub-type**	**Method**
Network-based methods	Divisive methods *Divides the PPIN into smaller networks and then search for protein complexes*	Friedel et al. [100]
		Pu et al. [101]
		MCL [102]
	Agglomerative methods *Takes the PPIN as a whole and then tries to find the complexes from it using various approaches*	ClusterONE [103]
		HACO [104]
		CMC [105]
		LCMA [106]
		MCODE [107]
Biological-context-aware methods	Core-attachment structure *Tries to find the core proteins and then add other proteins that are connected to it.*	DCA [108]
		MCL-Caw [109]
		CACHET [110]
		CORE [111]
		COACH [112]
	Functional information based	Zhang et al. [113]

	Adds information of protein functions to find the proteins which have same or similar functions.	CFOCM [114]
		DyCluster [115]
		PCP [116]
		DECAFF [117]
		RNSC [118]
	Incorporating evolutionary information *Adds information on the proteins from various organisms; orthologs, paralogs, gene fusion, etc.*	UEDAMAlign [119]
		COCIN [120]
		Hirsh et al. [121]
		Sharan et al. [122]
		Sharan et al. [123]
		Kelley et al. [124]
Specialized Methods	Sparse complexes *Using PPIN, dense interactions are taken and complexes are searched however, some complexes are spared and need different approaches to identify them.*	SWC [125]
		SPARC [126]
	Small complexes *To find the protein complexes of size three or two.*	CPredictor2.0 [127]
		CPredictor3.0 [128]
		SSS [129]
		Ruan et al. [130]

Many databases for protein complexes exist, however, most of them are either non-functional or have not been updated in at least the last two years. Nevertheless, there are two manually annotated and regularly updated protein complex databases, namely CORUM [131] and Complex portal [132]. Both define

the protein complexes based on various sets of evidence. For instance, CORUM, lists five levels of evidence that is based on the quality of confidence: experimental evidence, evidence from literature like reviews, known mammalian homolog, high-throughput experiment, and predicted function. Similarly, Complex portal uses six evidence levels: physical interaction evidence (ECO:0000353), experimental evidence from mixed species (ECO:0005543), homology evidence (ECO:0005610), orthology evidence (ECO:0005544), paralogy evidence (ECO:0005546), and inference from background scientific knowledge (ECO:0005547); given with their corresponding evidence and conclusion ontology [133]. Both databases also provide information on the stoichiometry of the complexes.

The subunits of a protein complex and their respective stoichiometry define one aspect of the protein complexes, the assembly of protein complexes is the other. The identification of the order of assembly is currently done using electrospray mass spectrometry (mass spectrometry using electrospray ionization), which can identify different intermediate states simultaneously. Through the knowledge gained from these studies, it has been discovered that most complexes follow an ordered assembly pathway [134]. Though highly regulated, a misassembled or disordered assembly can happen and may lead to disastrous consequences [135], such as aggregation and protein complex dysfunction [136]. Disordered assembly can occur due to, but not limited to, unfavorable environmental conditions such as unfavorable pH or temperature, mutations in genes for the protein subunits, and error in the folding process.

Aberrant changes in the pH and temperature of the body affect the folding of the proteins and under extreme conditions, protein denaturation can also occur. The deviant protein structure influences the interactions between different proteins and hence the formation of the complexes. Among the different attraction interactions, pH affects notably the salt bridges and hydrogen bonding. Disruption of these forces results in unfolded, misfolded, or differently folded protein. Nitrophorin 4 (NP4) protein releases nitric oxide (NO) in a pH-sensitive manner. At pH 5.5 NP4 is in a closed conformation where NO is tightly bound, while at pH 7.5 Asp30 becomes deprotonated, causing the conformation to change to an open state from which NO can easily escape [137]. However, it has been demonstrated that many macromolecules can tolerate small pH fluctuations that are inevitable with cellular functions [138].

Many misfolded proteins involved in diseases contain one or more mutations that destabilize the correct fold and/or stabilize a misfolded state [139]. When forming complexes, such mutations can result in perturbed inter-subunit interactions, dominant-negative mechanism, and stoichiometric imbalances [95]. Perturbation of the inter-subunit interaction can be witnessed in the case of acetyl-CoA dehydrogenase that causes the metabolic disorder. A missense mutation in the homo-tetrameric medium-chain of acetyl-CoA dehydrogenase results in the disruption of ionic bonds and water bridges. The substitution of a positively charged Lys to the negatively charged Glu at the homomeric interface affects the assembly of a functioning tetramer [140]. The dominant-negative mechanism arises when there is a mutation in one allele of a protein-coding gene and the complex is formed from the mix of wild and mutant proteins and the function is obstructed by the mutant protein [95]. The dominant-negative effect has been observed in STAT3 that belongs to the signal transducer and activator of transcription family. In the DNA binding domain of STAT3, five distinct mutations were noted. The STAT3 protein expression was unperturbed, phosphorylation, and dimerization capability were unchanged from the wild type. However, the STAT3 dimers that contained at least one mutant monomer presented severely reduced DNA-binding ability and resulted in hyper-IgE syndrome [141]. Stoichiometric imbalances can be observed in heteromeric complexes when one of the subunits is increased or decreased. Such imbalances are caused by either heterozygous mutations in individual protein-coding genes that decrease or knock out the expression of the allele or from larger scale copy number variations or aneuploidy [142].

Lastly, the protein folding machinery, though intricate, can still make errors, resulting in misfolded proteins, hence affecting the complex formation. Recently, it has been illustrated that subunits of many protein complexes start assembling before all the subunits are formed [143, 144]. The protein complex assembly occurs co-translationally and translation, folding, and assembly of protein complexes are integrated processes in eukaryotes [143]. The information of co-translational assembly of the proteins can be used to model an *in silico* approach taking the transcriptomics data across several time points. Emphasizing that the expression of the transcripts for the respective protein subunits in the complex peak when they are required to bind with the other subunits in course of protein complex formation.

The idea behind this thesis is that using the transcript expression would generate better inferences from RNA-Seq data. To validate this hypothesis we identified gaps in gene expression analyses and developed approaches to study transcript expression data. Furthermore, novel applications of the transcript expression are presented: assessing functional changes, identifying of biomarkers, and studying protein complexes. The novel approaches developed in this thesis were applied to liver transcriptomics data because the liver is involved in numerous functions and is prone to several diseases.

Liver – the organ of choice

To develop novel data analyzing methods aiming at proposing a solution for all these gaps identified in the current state of the art procedures, a single organ was chosen – the liver. The liver was selected owing to multiple reasons, such as the number of various functions it is involved in (discussed below), its significance in toxicological and drug safety studies, the amount of transcriptomics dataset available in the public domain, and the data generated in the EU-ToxRisk project (c.f. chapter 2). The liver plays a vital role in numerous functions in the body, ranging from digestion to metabolism and storage of various biomolecules to removing toxins from the body to maintaining homeostasis.

The liver plays a significant role in digestion, mainly of fats and carbohydrates. Bile, a dark-green-to-yellowish-brown fluid that is secreted from the liver and then stored and concentrated in the gall bladder, which is a key element in the digestion of fats. During digestion in the small intestine, bile is released into the duodenum where it breaks down the fats for further digestion and absorption [145]. It also helps in the absorption of fat-soluble vitamins such as A, D, E, and K [146]. The liver is also involved in controlling carbohydrate metabolism and maintains glucose concentrations in the normal range by tightly regulating the system of enzymes and kinases. The liver controls both, glucose breakdown and synthesis in hepatocytes. This process is under the control of glucoregulatory mediators among which insulin plays a key role [147].

The liver plays an equally important part in detoxification that it accomplishes by breaking down and removing harmful substances. It neutralizes a wide range of toxic chemicals, produced within the body or coming from the environment. These include, but are not limited to, alcohol [148], drugs [149], bile salts [150], and

bilirubin [151]. The liver achieves the task of detoxification of the system in different ways, as listed: removing large toxins from the blood, excreting cholesterol and other fat-soluble toxins in the bile, and finally breaking down and neutralizing the toxins. The breakdown is achieved in three steps termed as phase I, II, and III metabolism. Phase I involves functionalization reactions, phase II is a conjugation reaction, and phase III refers to the transporter-mediated elimination of drugs and/or metabolites from the body normally via liver, gut, kidney, or lung [149].

The liver is also an important immune tissue and is a part of the mononuclear phagocyte system. It houses the largest population of resident tissue macrophages in the body, Kupffer cells, which are responsible for phagocytosis [152]. The liver is capable of detecting the pathogens such as bacteria, viruses, and macromolecules entering the body via the gut. The liver aims at establishing a balance between being anti-inflammatory or immunotolerant (to allow food particles to pass) and exhibiting an immune response (to detect, capture, and eliminate the harmful pathogens). In the absence of infection, excessive inflammation may lead to sterile liver injury, tissue damage, and remodeling. On the other hand, insufficient immunity can result in chronic infection and cancer [153].

Liver stores and regulates the level of glycogen, iron, copper, and vitamin A in the body. After a meal, when there is excess sugar in the blood, the liver removes it from the system via the blood in the portal vein and stores it in the form of glycogen. When blood sugar levels fall, the stored glycogen is released back into the system by breaking it down as glucose [154]. The excess iron released from the breakdown of the red blood cells (RBCs) is stored in the hepatocytes and the liver acts as a regulator of iron homeostasis in the body as well [155]. Similarly, copper is also stored in the liver and it plays a protective role against copper-induced cytotoxicity [156]. The liver also stores a significant amount of vitamin A. From the circulation, vitamin A is taken up by the hepatocytes and ~80 % of it is stored in hepatic stellate cells [157].

Moreover, the liver is the body's chemical "factory." It can synthesize several chemicals needed by the body to function, using the raw materials absorbed by the intestine. Various chemicals synthesized by the liver are listed ahead with their functions. Albumin is required to transport fatty acids and steroid hormones as well as maintaining oncotic pressure [158]. Angiotensinogen is a component of the renin-angiotensin system (RAS) that regulates blood pressure and fluid balance

[159]. Coagulation factors help in achieving hemostasis [160]. Complement factors are part of the complement system [161]. Haptoglobin binds to free plasma hemoglobin and allows degradative enzymes to destroy it [162]. Caeruloplasmin is responsible for oxidation of Fe^{2+} into Fe^{3+} [163]. Lastly, transferrin binds to iron and transports it throughout the body [164].

Owing to the multitude of functions that the liver is involved in, it is continuously exposed to various toxic substances from within the body and outside (gut). This makes the liver highly susceptible to injury, infection, and diseases. One of the most common liver diseases is hepatitis, which is usually caused by viruses like hepatitis A, B, and C [165, 166]. Hepatitis can also be caused by non-infectious causes including, drugs, allergic reactions, heavy drinking, or obesity [167-169]. Long-term liver damage from any causes can result in permanent scarring and loss of function, such a condition is referred to as cirrhosis [170]. The liver is also prone to various types of cancer, namely hepatocellular carcinoma (HCC) [171], intrahepatic cholangiocarcinoma (bile duct cancer) [172], angiosarcoma (cancers beginning in cells lining the blood vessels of the liver) [173], and hepatoblastoma (liver cancer in children) [174]. Of all, HCC is the most common liver cancer and accounts for ~80% of liver cancers [175]. Liver cancers usually occur after long-term hepatitis B or C infection and/or cirrhosis [176, 177]. Some other less frequently occurring liver diseases include ascites (leakage of liver fluid into the belly) [178], gallstones (stuck gallstone in the bile duct) [179], hemochromatosis (high iron deposit in the liver) [180], primary sclerosing cholangitis (inflammation and scarring in the bile ducts) [181], primary biliary cirrhosis (slowly destroys the bile ducts) [182], and liver failure [183].

Aims, objective, and outline of the thesis

Several liver cell models ranging from cancer cell models to iPSC derived cell models to liver cancer slices were made available through the EU-ToxRisk project were sequenced in-house. All cell models were at baseline. An investigation to find which cell model closely mimics the liver biopsies, an exhaustive comparison of these cell models was done using transcriptomics data. Besides finding the most similar liver cell model to *in vivo* liver, the differences in the gene-based and transcript-based analyses were highlighted. In **Chapter 2**, the similarity of the liver cell models to *in vivo* liver and the advantages of using transcript expression over gene expression

analyses are discussed in detail. Liver comparison results are provided as differentially expressed genes, differential transcript usage, and their coverage on human KEGG pathways.

From the knowledge gained in the previous chapter on the transcripts, especially the protein coding transcripts, it was realized that there are a significant number of protein coding transcripts, originating from the same or different genes, that code for similar proteins. Coding for similar protein would imply a similar function, and if such transcripts were studied individually, the functional characteristics might be masked. To address the concern of different transcripts that code for the same protein, FuSe (Functional grouping of transcripts for RNA-Seq analyses) was developed. **Chapter 3** discusses the development of FuSe and its application on liver transcriptomics data. The transcriptomics data was obtained for liver cell models exposed to toxic and therapeutic APAP doses, along with untreated and DMSO control.

Furthermore, in **Chapter 4**, the ability of machine learning (ML) in finding novel biomarkers using transcriptomics data is investigated. Using various machine learning algorithms and approaches, a novel methodology to search for better disease biomarkers is developed. To illustrate the applicability of the methodology, transcriptomics data for several cell models of hepatocellular carcinoma (HCC) and healthy liver biopsies were taken. For the identification of the biomarkers, transcript expression rather than gene expression was used and it is illustrated that transcript biomarkers give high accuracy in differentiating the healthy and HCC cell models.

Through transcript expression comparison a list of perturbed transcripts is obtained, however, the knowledge of their binding partners can further aid in better understanding the changes in the biological system. These binding partners might form protein complexes and perform specific functions. In **Chapter 5**, time-series transcriptomics data from untreated liver cell models were taken to first devise an approach to calculate protein complex expression and then to predict the order of assembly of the protein complexes. Along with transcript expression data, stoichiometry was used to calculate the protein complex expression and for the prediction of the order of assembly, dynamic Bayesian networks were used. The challenges in achieving the desired output and ways to overcome them are presented.

In **Chapter 6**, a discussion on the overall conclusions derived from this thesis is presented and further ideas to explore are debated. Starting with liver cell model comparison, illustrating the exhaustive pathway mapping data generated for several cell models and its applicability in selecting a cell model for a study. Furthermore, the advantages of using transcript expression over gene expression for transcriptomics data analyses are discussed. Then moving to the tool, namely FuSe, which is developed to analyze mRNA transcripts that code for similar proteins. The opportunities for the improvement of FuSe and the addition of new features like non-coding transcripts are presented. Furthermore, the interesting findings from the application of ML to transcriptomics data in the quest of finding novel transcript biomarkers in HCC cell models are given. Its intended use for other diseases and on non-invasive transcriptomics data is discussed. Finally, results from the protein complex expression and assembly order prediction are discussed.

Lastly, under impact, a scrutiny of the outcomes from chapters 2 to 5 are presented with arguments on its impact and practicality.

References

1. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors**. *Proceedings of the National Academy of Sciences* 1977, **74**(12):5463-5467.
2. Mardis ER: **Next-generation DNA sequencing methods**. *Annu Rev Genomics Hum Genet* 2008, **9**:387-402.
3. Heather JM, Chain B: **The sequence of sequencers: The history of sequencing DNA**. *Genomics* 2016, **107**(1):1-8.
4. Kulski JK: **Next-generation sequencing—an overview of the history, tools, and "Omic" applications**. *Next Generation Sequencing–Advances, Applications and Challenges* 2016:3-60.
5. Ebili HO, Hassall JC, Fadhil W, Ham-Karim H, Asiri A, Raposo TP, Agboola AJ, Ilyas M: **"Squirrel" Primer-Based PCR Assay for Direct and Targeted Sanger Sequencing of Short Genomic Segments**. *J Biomol Tech* 2017, **28**(3):97-110.
6. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE: **Big data: astronomical or genomical?** *PLoS biology* 2015, **13**(7):e1002195.
7. Beadle GW, Tatum EL: **Genetic Control of Biochemical Reactions in Neurospora**. *Proc Natl Acad Sci U S A* 1941, **27**(11):499-506.
8. The_UniProt_Consortium: **UniProt: a hub for protein information**. *Nucleic Acids Res* 2015, **43**(Database issue):27.
9. Frankish A, Vullo A, Zadissa A, Yates A, Thormann A, Parker A, Gall A, Moore B, Walts B, Aken BL *et al*: **Ensembl 2018**. *Nucleic Acids Res* 2017, **46**(D1):D754-D761.

10. Roth MJ, Forbes AJ, Boyne MT, 2nd, Kim YB, Robinson DE, Kelleher NL: **Precise and parallel characterization of coding polymorphisms, alternative splicing, and modifications in human proteins by mass spectrometry**. *Mol Cell Proteomics* 2005, **4**(7):1002-1008.
11. Karlsson C, Malmström L, Aebersold R, Malmström J: **Proteome-wide selected reaction monitoring assays for the human pathogen Streptococcus pyogenes**. *Nat Commun* 2012, **3**(1301).
12. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic Acids Res* 2007, **35**(Database issue):27.
13. He F, Li X, Spatrick P, Casillo R, Dong S, Jacobson A: **Genome-wide analysis of mRNAs regulated by the nonsense-mediated and 5' to 3' mRNA decay pathways in yeast**. *Mol Cell* 2003, **12**(6):1439-1452.
14. Celik A, Kervestin S, Jacobson A: **NMD: At the crossroads between translation termination and ribosome recycling**. *Biochimie* 2015, **114**:2-9.
15. Pawlicka K, Kalathiya U, Alfaro J: **Nonsense-Mediated mRNA Decay: Pathologies and the Potential for Novel Therapeutics**. *Cancers* 2020, **12**(3):765.
16. Lykke-Andersen S, Jensen TH: **Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes**. *Nat Rev Mol Cell Biol* 2015, **16**(11):665-677.
17. Klauer AA, van Hoof A: **Degradation of mRNAs that lack a stop codon: a decade of nonstop progress**. *Wiley Interdiscip Rev RNA* 2012, **3**(5):649-660.
18. Pelechano V, Steinmetz LM: **Gene regulation by antisense transcription**. *Nat Rev Genet* 2013, **14**(12):880-893.
19. Ransohoff JD, Wei Y, Khavari PA: **The functions and unique features of long intergenic non-coding RNA**. *Nat Rev Mol Cell Biol* 2018, **19**(3):143-157.
20. Guenzl PM, Barlow DP: **Macro lncRNAs: a new layer of cis-regulatory information in the mammalian genome**. *RNA Biol* 2012, **9**(6):731-741.
21. Bernardo BC, Ooi JY, Lin RC, McMullen JR: **miRNA therapeutics: a new class of drugs with potential therapeutic applications in the heart**. *Future Med Chem* 2015, **7**(13):1771-1792.
22. Vishnoi A, Rani S: **MiRNA Biogenesis and Regulation of Diseases: An Overview**. *Methods Mol Biol* 2017:6524-6523_6521.
23. Krauskopf J, Caiment F, van Veldhoven K, Chadeau-Hyam M, Sinharay R, Chung KF, Cullinan P, Collins P, Barratt B, Kelly FJ *et al*: **The human circulating miRNome reflects multiple organ disease risks in association with short-term exposure to traffic-related air pollution**. *Environ Int* 2018, **113**:26-34.
24. Czech B, Munafò M, Ciabrelli F, Eastwood EL, Fabry MH, Kneuss E, Hannon GJ: **piRNA-Guided Genome Defense: From Biogenesis to Silencing**. *Annu Rev Genet* 2018, **52**:131-157.
25. Nikam RR, Gore KR: **Journey of siRNA: Clinical Developments and Targeted Delivery**. *Nucleic Acid Ther* 2018, **28**(4):209-224.
26. Laganà A, Veneziano D, Russo F, Pulvirenti A, Giugno R, Croce CM, Ferro A: **Computational design of artificial RNA molecules for gene regulation**. *Methods Mol Biol* 2015:2291-2298_2225.
27. Will CL, Lührmann R: **Spliceosome structure and function**. *Cold Spring Harb Perspect Biol* 2011, **3**(7).
28. Maden BE, Hughes JM: **Eukaryotic ribosomal RNA: the recent excitement in the nucleotide modification problem**. *Chromosoma* 1997, **105**(7-8):391-400.
29. Huang C, Shi J, Guo Y, Huang W, Huang S, Ming S, Wu X, Zhang R, Ding J, Zhao W: **A snoRNA modulates mRNA 3' end processing and regulates the expression of a subset of mRNAs**. *Nucleic Acids Res* 2017, **45**(15):8647-8660.
30. Mossink MH, van Zon A, Scheper RJ, Sonneveld P, Wiemer EAC: **Vaults: a ribonucleoprotein particle involved in drug resistance?** *Oncogene* 2003, **22**(47):7458-7467.

31. Büscher M, Horos R, Hentze MW: **'High vault-age': non-coding RNA control of autophagy**. *Open Biology* 2020, **10**(2):190307.
32. Xiao Y-S, Zeng D, Liang Y-K, Wu Y, Li M-F, Qi Y-Z, Wei X-L, Huang W-H, Chen M, Zhang G-J: **Major vault protein is a direct target of Notch1 signaling and contributes to chemoresistance in triple-negative breast cancer cells**. *Cancer Letters* 2019, **440-441**:156-167.
33. Yu S, Lemos B: **A Portrait of Ribosomal DNA Contacts with Hi-C Reveals 5S and 45S rDNA Anchoring Points in the Folded Human Genome**. *Genome Biol Evol* 2016, **8**(11):3545-3558.
34. Zhao J, Qin B, Nikolay R, Spahn CMT, Zhang G: **Translatomics: The Global View of Translation**. *International journal of molecular sciences* 2019, **20**(1):212.
35. Wade JT: **Where to begin? Mapping transcription start sites genome-wide in Escherichia coli**. *Journal of bacteriology* 2015, **197**(1):4-6.
36. Licatalosi DD, Darnell RB: **RNA processing and its regulation: global insights into biological networks**. *Nat Rev Genet* 2010, **11**(1):75-87.
37. Rojas-Duran MF, Gilbert WV: **Alternative transcription start site selection leads to large differences in translation activity in yeast**. *RNA (New York, NY)* 2012, **18**(12):2299-2305.
38. Zhang Y, Liu XS, Liu QR, Wei L: **Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species**. *Nucleic Acids Res* 2006, **34**(12):3465-3475.
39. Peters BA, St Croix B, Sjöblom T, Cummins JM, Silliman N, Ptak J, Saha S, Kinzler KW, Hatzis C, Velculescu VE: **Large-scale identification of novel transcripts in the human genome**. *Genome Res* 2007, **17**(3):287-292.
40. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: **The KEGG resource for deciphering the genome**. *Nucleic Acids Res* 2004, **32**(suppl_1):D277-D280.
41. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B: **The reactome pathway knowledgebase**. *Nucleic Acids Res* 2018, **46**(D1):D649-D655.
42. Karp PD, Billington R, Caspi R, Fulcher CA, Latendresse M, Kothari A, Keseler IM, Krummenacker M, Midford PE, Ong Q: **The BioCyc collection of microbial genomes and metabolic pathways**. *Briefings in bioinformatics* 2019, **20**(4):1085-1093.
43. The Gene Ontology Consortium: **The Gene Ontology Resource: 20 years and still GOing strong**. *Nucleic Acids Res* 2018, **47**(D1):D330-D338.
44. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI**. *Nucleic Acids Res* 2005, **33**(suppl_1):D54-D58.
45. Sakuratani Y, Horie M, Leinala E: **Integrated Approaches to Testing and Assessment: OECD Activities on the Development and Use of Adverse Outcome Pathways and Case Studies**. *Basic & Clinical Pharmacology & Toxicology* 2018, **123**(S5):20-28.
46. Smedley D, Haider S, Durinck S, Pandini L, Provero P, Allen J, Arnaiz O, Awedh MH, Baldock R, Barbiera G *et al*: **The BioMart community portal: an innovative alternative to large, centralized data repositories**. *Nucleic Acids Res* 2015, **43**(W1):20.
47. Rodriguez JM, Rodriguez-Rivas J, Di Domenico T, Vázquez J, Valencia A, Tress ML: **APPRIS 2017: principal isoforms for multiple gene sets**. *Nucleic Acids Res* 2017, **46**(D1):D213-D217.
48. Carninci P: **Is sequencing enlightenment ending the dark age of the transcriptome?** *Nat Methods* 2009, **6**(10):711-713.
49. Lopes R, Agami R, Korkmaz G: **GRO-seq, A Tool for Identification of Transcripts Regulating Gene Expression**. *Methods Mol Biol* 2017:6716-6712_6713.

50. Hu X, Wang Q, Tang M, Barthel F, Amin S, Yoshihara K, Lang FM, Martinez-Ledesma E, Lee SH, Zheng S *et al*: **TumorFusions: an integrative resource for cancer-associated transcript fusions**. *Nucleic Acids Res* 2018, **46**(D1):D1144-D1149.
51. Lee YS, Won KH, Oh JD, Shin D: **In silico approach to calculate the transcript capacity**. *Genomics Inform* 2019, **17**(3):26.
52. Schroeder DI, Myers RM: **Multiple transcription start sites for FOXP2 with varying cellular specificities**. *Gene* 2008, **413**(1-2):42-48.
53. Reyes A, Huber W: **Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues**. *Nucleic Acids Res* 2018, **46**(2):582-592.
54. Kamieniarz-Gdula K, Proudfoot NJ: **Transcriptional Control by Premature Termination: A Forgotten Mechanism**. *Trends Genet* 2019, **35**(8):553-564.
55. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic Acids Res* 2000, **28**(1):235-242.
56. Kandathil SM, Greener JG, Jones DT: **Recent developments in deep learning applied to protein structure prediction**. *Proteins: Structure, Function, and Bioinformatics* 2019, **87**(12):1179-1189.
57. Jeon J, Nim S, Teyra J, Datti A, Wrana JL, Sidhu SS, Moffat J, Kim PM: **A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening**. *Genome Med* 2014, **6**(7):014-0057.
58. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, Li B, Madabhushi A, Shah P, Spitzer M *et al*: **Applications of machine learning in drug discovery and development**. *Nature Reviews Drug Discovery* 2019, **18**(6):463-477.
59. Olivecrona M, Blaschke T, Engkvist O, Chen H: **Molecular de-novo design through deep reinforcement learning**. *J Cheminform* 2017, **9**(1):017-0235.
60. Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, Bansal M, Ammad-ud-din M, Hintsanen P, Khan SA *et al*: **A community effort to assess and improve drug sensitivity prediction algorithms**. *Nat Biotechnol* 2014, **32**(12):1202-1212.
61. Leung MK, Xiong HY, Lee LJ, Frey BJ: **Deep learning of the tissue-regulated splicing code**. *Bioinformatics* 2014, **30**(12).
62. Xia X, Chen W, McDermott J, Han JJ: **Molecular and phenotypic biomarkers of aging**. *F1000Res* 2017, **6**(860).
63. Vernon ST, Hansen T, Kott KA, Yang JY, O'Sullivan JF, Figtree GA: **Utilizing state-of-the-art "omics" technology and bioinformatics to identify new biological mechanisms and biomarkers for coronary artery disease**. *Microcirculation* 2019, **26**(2):23.
64. Opal SM, Wittebole X: **Biomarkers of Infection and Sepsis**. *Crit Care Clin* 2020, **36**(1):11-22.
65. Tolios A, De Las Rivas J, Hovig E, Trouillas P, Scorilas A, Mohr T: **Computational approaches in cancer multidrug resistance research: Identification of potential biomarkers, drug targets and drug-target interactions**. *Drug Resist Updat* 2020, **48**(100662):18.
66. Tolios A, De Las Rivas J, Hovig E, Trouillas P, Scorilas A, Mohr T: **Computational approaches in cancer multidrug resistance research: Identification of potential biomarkers, drug targets and drug-target interactions**. *Drug Resistance Updates* 2020, **48**:100662.
67. Strimbu K, Tavel JA: **What are biomarkers?** *Curr Opin HIV AIDS* 2010, **5**(6):463-466.
68. Davis MA, Eldridge S, Louden C: **Chapter 10 - Biomarkers: Discovery, Qualification and Application**. In: *Haschek and Rousseaux's Handbook of Toxicologic Pathology (Third Edition).* Edited by Haschek WM, Rousseaux CG, Wallig MA. Boston: Academic Press; 2013: 317-352.

69. Huss R: **Chapter 19 - Biomarkers**. In: *Translational Regenerative Medicine.* Edited by Atala A, Allickson JG. Boston: Academic Press; 2015: 235-241.
70. Mayeux R: **Biomarkers: potential uses and limitations**. *NeuroRx* 2004, **1**(2):182-188.
71. Henry NL, Hayes DF: **Cancer biomarkers**. *Molecular oncology* 2012, **6**(2):140-146.
72. Brody T: **Chapter 19 - Biomarkers**. In: *Clinical Trials (Second Edition).* Edited by Brody T. Boston: Academic Press; 2016: 377-419.
73. Pinkerton L, Hein M, Stayner L: **Mortality among a cohort of garment workers exposed to formaldehyde: an update**. *Occupational and Environmental Medicine* 2004, **61**(3):193-200.
74. Costa S, Coelho P, Costa C, Silva S, Mayan O, Santos LS, Gaspar J, Teixeira JP: **Genotoxic damage in pathology anatomy laboratory workers exposed to formaldehyde**. *Toxicology* 2008, **252**(1-3):40-48.
75. Costa S, Costa C, Madureira J, Valdiglesias V, Teixeira-Gomes A, de Pinho PG, Laffon B, Teixeira JP: **Occupational exposure to formaldehyde and early biomarkers of cancer risk, immunotoxicity and susceptibility**. *Environmental research* 2019, **179**:108740.
76. Gonçalves MdS, Moura Neto JPd, Souza C, Melo P, Reis MGd: **Evaluating glutathione S-Transferase (GST) null genotypes (GSTT1 and GSTM1) as a potential biomarker of predisposition for developing leukopenia**. *International Journal of Laboratory Hematology* 2010, **32**(1p1):e49-e56.
77. Long NP, Park S, Anh NH, Nghi TD, Yoon SJ, Park JH, Lim J, Kwon SW: **High-throughput Omics and statistical learning integration for the discovery and validation of novel diagnostic signatures in colorectal cancer**. *International journal of molecular sciences* 2019, **20**(2):296.
78. Kaur H, Dhall A, Kumar R, Raghava GPS: **Identification of Platform-Independent Diagnostic Biomarker Panel for Hepatocellular Carcinoma Using Large-Scale Transcriptomics Data**. *Front Genet* 2020, **10**(1306).
79. van Ijzendoorn DGP, Szuhai K, Briaire-de Bruijn IH, Kostine M, Kuijjer ML, Bovée JVMG: **Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas**. *PLoS Comput Biol* 2019, **15**(2):e1006826-e1006826.
80. Mamoshina P, Volosnikova M, Ozerov IV, Putin E, Skibina E, Cortese F, Zhavoronkov A: **Machine Learning on Human Muscle Transcriptomic Data for Biomarker Discovery and Tissue-Specific Drug Target Identification**. *Front Genet* 2018, **9**(242).
81. Ding D, Han S, Zhang H, He Y, Li Y: **Predictive biomarkers of colorectal cancer**. *Comput Biol Chem* 2019, **83**(107106):3.
82. Yang Z, Liang X, Fu Y, Liu Y, Zheng L, Liu F, Li T, Yin X, Qiao X, Xu X: **Identification of AUNIP as a candidate diagnostic and prognostic biomarker for oral squamous cell carcinoma**. *EBioMedicine* 2019, **47**:44-57.
83. Long NP, Park S, Anh NH, Nghi TD, Yoon SJ, Park JH, Lim J, Kwon SW: **High-Throughput Omics and Statistical Learning Integration for the Discovery and Validation of Novel Diagnostic Signatures in Colorectal Cancer**. *Int J Mol Sci* 2019, **20**(2).
84. Szostaczuk N, van Schothorst EM, Sánchez J, Priego T, Palou M, Bekkenkamp-Grovenstein M, Faustmann G, Obermayer-Pietsch B, Tiran B, Roob JM: **Identification of blood cell transcriptome-based biomarkers in adulthood predictive of increased risk to develop metabolic disorders using early life intervention rat models**. *The FASEB Journal* 2020.
85. Delpuech O, Rooney C, Mooney L, Baker D, Shaw R, Dymond M, Wang D, Zhang P, Cross S, Veldman-Jones M *et al*: **Identification of Pharmacodynamic Transcript Biomarkers in Response to FGFR Inhibition by AZD4547**. *Mol Cancer Ther* 2016, **15**(11):2802-2813.
86. Dwane S, Kiely PA: **Tools used to study how protein complexes are assembled in signaling cascades**. *Bioeng Bugs* 2011, **2**(5):247-259.

87. Ellis RJ: **Macromolecular crowding: an important but neglected aspect of the intracellular environment**. *Current Opinion in Structural Biology* 2001, **11**(1):114-119.
88. McGuffee SR, Elcock AH: **Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm**. *PLoS Comput Biol* 2010, **6**(3):e1000694.
89. Perkins JR, Diboun I, Dessailly BH, Lees JG, Orengo C: **Transient protein-protein interactions: structural, functional, and network properties**. *Structure* 2010, **18**(10):1233-1243.
90. Marsh JA, Teichmann SA: **Structure, Dynamics, Assembly, and Evolution of Protein Complexes**. *Annual review of biochemistry* 2015, **84**(1):551-575.
91. Perkins JR, Diboun I, Dessailly BH, Lees JG, Orengo C: **Transient protein-protein interactions: structural, functional, and network properties**. *Structure* 2010, **18**(10):1233-1243.
92. Tompa P, Davey NE, Gibson TJ, Babu MM: **A million peptide motifs for the molecular biologist**. *Molecular cell* 2014, **55**(2):161-169.
93. Levy ED, De S, Teichmann SA: **Cellular crowding imposes global constraints on the chemistry and evolution of proteomes**. *Proceedings of the National Academy of Sciences* 2012, **109**(50):20461-20466.
94. Landry CR, Levy ED, Abd Rabbo D, Tarassov K, Michnick SW: **Extracting insight from noisy cellular networks**. *Cell* 2013, **155**(5):983-989.
95. Bergendahl LT, Gerasimavicius L, Miles J, Macdonald L, Wells JN, Welburn JP, Marsh JA: **The role of protein complexes in human genetic disease**. *Protein Science* 2019, **28**(8):1400-1411.
96. Hashimoto K, Nishi H, Bryant S, Panchenko AR: **Caught in self-interaction: evolutionary and functional mechanisms of protein homooligomerization**. *Phys Biol* 2011, **8**(3):035007-035007.
97. Marsh JA, Teichmann SA: **Structure, dynamics, assembly, and evolution of protein complexes**. *Annual review of biochemistry* 2015, **84**:551-575.
98. Perica T, Marsh JA, Sousa FL, Natan E, Colwell LJ, Ahnert SE, Teichmann SA: **The emergence of protein complexes: quaternary structure, dynamics and allostery**. In.: Portland Press Ltd.; 2012.
99. Zahiri J, Emamjomeh A, Bagheri S, Ivazeh A, Mahdevar G, Tehrani HS, Mirzaie M, Fakheri BA, Mohammad-Noori M: **Protein complex prediction: A survey**. *Genomics* 2019.
100. Friedel CC, Krumsiek J, Zimmer R: **Bootstrapping the interactome: unsupervised identification of protein complexes in yeast**. In: *Annual International Conference on Research in Computational Molecular Biology: 2008*. Springer: 3-16.
101. Pu S, Vlasblom J, Emili A, Greenblatt J, Wodak SJ: **Identifying functional modules in the physical interactome of Saccharomyces cerevisiae**. *Proteomics* 2007, **7**(6):944-960.
102. Van Dongen SM: **Graph clustering by flow simulation**. 2000.
103. Nepusz T, Yu H, Paccanaro A: **Detecting overlapping protein complexes in protein-protein interaction networks**. *Nat Methods* 2012, **9**(5):471.
104. Wang H, Kakaradov B, Collins SR, Karotki L, Fiedler D, Shales M, Shokat KM, Walther TC, Krogan NJ, Koller D: **A complex-based reconstruction of the Saccharomyces cerevisiae interactome**. *Molecular & Cellular Proteomics* 2009, **8**(6):1361-1381.
105. Liu G, Wong L, Chua HN: **Complex discovery from weighted PPI networks**. *Bioinformatics* 2009, **25**(15):1891-1897.
106. Li X-L, Foo C-S, Tan S-H, Ng S-K: **Interaction graph mining for protein complexes using local clique merging**. *Genome informatics* 2005, **16**(2):260-269.
107. Bader GD, Hogue CW: **An automated method for finding molecular complexes in large protein interaction networks**. *BMC Bioinformatics* 2003, **4**(1):2.
108. Shen X, Yi L, Jiang X, He T, Yang J, Xie W, Hu P, Hu X: **Identifying protein complex by integrating characteristic of core-attachment into dynamic PPI network**. *PloS one* 2017, **12**(10):e0186134.

109. Srihari S, Ning K, Leong HW: **MCL-CAw: a refinement of MCL for detecting yeast complexes from weighted PPI networks by incorporating core-attachment structure**. *BMC Bioinformatics* 2010, **11**(1):504.

110. Wu M, Li X-l, Kwoh C-K, Ng S-K, Wong L: **Discovery of protein complexes with core-attachment structures from tandem affinity purification (TAP) data**. *Journal of Computational Biology* 2012, **19**(9):1027-1042.

111. Leung HC, Xiang Q, Yiu S-M, Chin FY: **Predicting protein complexes from PPI data: a core-attachment approach**. *Journal of Computational Biology* 2009, **16**(2):133-144.

112. Wu M, Li X, Kwoh C-K, Ng S-K: **A core-attachment based method to detect protein complexes in PPI networks**. *BMC Bioinformatics* 2009, **10**(1):169.

113. Zhang Z, Song J, Tang J, Xu X, Guo F: **Detecting complexes from edge-weighted PPI networks via genes expression analysis**. *BMC systems biology* 2018, **12**(4):40.

114. Li B, Liao B: **Protein complexes prediction method based on core—attachment structure and functional annotations**. *International journal of molecular sciences* 2017, **18**(9):1910.

115. Hanna EM, Zaki N, Amin A: **Detecting protein complexes in protein interaction networks modeled as gene expression biclusters**. *PloS one* 2015, **10**(12):e0144163.

116. Chua HN, Ning K, Sung W-K, Leong HW, Wong L: **Using indirect protein–protein interactions for protein complex prediction**. *Journal of bioinformatics and computational biology* 2008, **6**(03):435-466.

117. Li X-L, Foo C-S, Ng S-K: **Discovering protein complexes in dense reliable neighborhoods of protein interaction networks**. In: *Computational Systems Bioinformatics: (Volume 6).* World Scientific; 2007: 157-168.

118. King AD, Pržulj N, Jurisica I: **Protein complex prediction via cost-based clustering**. *Bioinformatics* 2004, **20**(17):3013-3020.

119. Peng W, Wang J, Wu F, Yi P: **Detecting conserved protein complexes using a dividing-and-matching algorithm and unequally lenient criteria for network comparison**. *Algorithms for Molecular Biology* 2015, **10**(1):21.

120. Nguyen P-V, Srihari S, Leong HW: **Identifying conserved protein complexes between species by constructing interolog networks**. *BMC Bioinformatics* 2013, **14**(S16):S8.

121. Hirsh E, Sharan R: **Identification of conserved protein complexes based on a model of protein network evolution**. *Bioinformatics* 2007, **23**(2):e170-e176.

122. Sharan R, Ideker T, Kelley B, Shamir R, Karp RM: **Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data**. *Journal of Computational Biology* 2005, **12**(6):835-846.

123. Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T: **Conserved patterns of protein interaction in multiple species**. *Proceedings of the National Academy of Sciences* 2005, **102**(6):1974-1979.

124. Kelley BP, Sharan R, Karp RM, Sittler T, Root DE, Stockwell BR, Ideker T: **Conserved pathways within bacteria and yeast as revealed by global protein network alignment**. *Proceedings of the National Academy of Sciences* 2003, **100**(20):11394-11399.

125. Yong CH, Liu G, Chua HN, Wong L: **Supervised maximum-likelihood weighting of composite protein networks for complex prediction**. In: *BMC systems biology: 2012*. Springer: S13.

126. Srihari S, Leong HW: **Employing functional interactions for characterisation and detection of sparse complexes from yeast PPI networks**. *International journal of bioinformatics research and applications* 2012, **8**(3-4):286-304.

127. Xu B, Wang Y, Wang Z, Zhou J, Zhou S, Guan J: **An effective approach to detecting both small and large complexes from protein-protein interaction networks**. *BMC Bioinformatics* 2017, **18**(12):419.
128. Xu Y, Zhou J, Zhou S, Guan J: **CPredictor3. 0: detecting protein complexes from PPI networks with expression data and functional annotations**. *BMC systems biology* 2017, **11**(7):45-56.
129. Yong CH, Maruyama O, Wong L: **Discovery of small protein complexes from PPI networks with size-specific supervised weighting**. *BMC systems biology* 2014, **8**(S5):S3.
130. Ruan P, Hayashida M, Maruyama O, Akutsu T: **Prediction of heterotrimeric protein complexes by two-phase learning using neighboring kernels**. In: *BMC Bioinformatics: 2014*. Springer: S6.
131. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes H-W: **CORUM: the comprehensive resource of mammalian protein complexes—2009**. *Nucleic Acids Res* 2009, **38**(suppl_1):D497-D501.
132. Meldal BHM, Forner-Martinez O, Costanzo MC, Dana J, Demeter J, Dumousseau M, Dwight SS, Gaulton A, Licata L, Melidoni AN *et al*: **The complex portal - an encyclopaedia of macromolecular complexes**. *Nucleic Acids Res* 2014, **43**(D1):D479-D484.
133. Giglio M, Tauber R, Nadendla S, Munro J, Olley D, Ball S, Mitraka E, Schriml LM, Gaudet P, Hobbs ET: **ECO, the Evidence & Conclusion Ontology: community standard for evidence information**. *Nucleic Acids Res* 2019, **47**(D1):D1186-D1194.
134. Marsh JA, Hernández H, Hall Z, Ahnert SE, Perica T, Robinson CV, Teichmann SA: **Protein complexes are under evolutionary selection to assemble via ordered pathways**. *Cell* 2013, **153**(2):461-470.
135. Dobson CM: **Protein folding and misfolding**. *Nature* 2003, **426**(6968):884-890.
136. Sudha G, Nussinov R, Srinivasan N: **An overview of recent advances in structural bioinformatics of protein-protein interactions and a guide to their principles**. *Prog Biophys Mol Biol* 2014, **116**(2-3):141-150.
137. Di Russo NV, Estrin DA, Martí MA, Roitberg AE: **pH-Dependent Conformational Changes in Proteins and Their Effect on Experimental pKas: The Case of Nitrophorin 4**. *PLoS Comput Biol* 2012, **8**(11):e1002761.
138. Talley K, Alexov E: **On the pH-optimum of activity and stability of proteins**. *Proteins* 2010, **78**(12):2699-2706.
139. Valastyan JS, Lindquist S: **Mechanisms of protein-folding diseases at a glance**. *Dis Model Mech* 2014, **7**(1):9-14.
140. Yokota I, Saijo T, Vockley J, Tanaka K: **Impaired tetramer assembly of variant medium-chain acyl-coenzyme A dehydrogenase with a glutamate or aspartate substitution for lysine 304 causing instability of the protein**. *Journal of Biological Chemistry* 1992, **267**(36):26004-26010.
141. Minegishi Y, Saito M, Tsuchiya S, Tsuge I, Takada H, Hara T, Kawamura N, Ariga T, Pasic S, Stojkovic O: **Dominant-negative mutations in the DNA-binding domain of STAT3 cause hyper-IgE syndrome**. *Nature* 2007, **448**(7157):1058-1062.
142. Rice AM, McLysaght A: **Dosage sensitivity is a major determinant of human copy number variant pathogenicity**. *Nature communications* 2017, **8**(1):1-11.
143. Shiber A, Döring K, Friedrich U, Klann K, Merker D, Zedan M, Tippmann F, Kramer G, Bukau B: **Cotranslational assembly of protein complexes in eukaryotes revealed by ribosome profiling**. *Nature* 2018, **561**(7722):268-272.
144. Mayr C: **Protein complexes assemble as they are being made**. In.: Nature Publishing Group; 2018.
145. Chen I, Cassaro S: **Physiology, Bile Acids**.
146. McMillin M, DeMorrow S: **Effects of bile acids on neurological function and disease**. *FASEB J* 2016, **30**(11):3658-3668.

147. Raddatz D, Ramadori G: **Carbohydrate metabolism and the liver: actual aspects from physiology and disease**. *Z Gastroenterol* 2007, **45**(1):51-62.
148. Rocco A, Compare D, Angrisani D, Sanduzzi Zamparelli M, Nardone G: **Alcoholic disease: liver and beyond**. *World J Gastroenterol* 2014, **20**(40):14652-14659.
149. Almazroo OA, Miah MK, Venkataramanan R: **Drug Metabolism in the Liver**. *Clin Liver Dis* 2017, **21**(1):1-20.
150. Ballatori N, Rebbeor JF, Connolly GC, Seward DJ, Lenth BE, Henson JH, Sundaram P, Boyer JL: **Bile salt excretion in skate liver is mediated by a functional analog of Bsep/Spgp, the bile salt export pump**. *Am J Physiol Gastrointest Liver Physiol* 2000, **278**(1).
151. Sullivan JI, Rockey DC: **Diagnosis and evaluation of hyperbilirubinemia**. *Curr Opin Gastroenterol* 2017, **33**(3):164-170.
152. Dixon LJ, Barnes M, Tang H, Pritchard MT, Nagy LE: **Kupffer cells in the liver**. *Compr Physiol* 2013, **3**(2):785-797.
153. Kubes P, Jenne C: **Immune Responses in the Liver**. *Annu Rev Immunol* 2018, **36**:247-277.
154. Han HS, Kang G, Kim JS, Choi BH, Koo SH: **Regulation of glucose metabolism from a liver-centric perspective**. *Exp Mol Med* 2016, **48**(3):122.
155. Rishi G, Subramaniam VN: **The liver in regulation of iron homeostasis**. *American Journal of Physiology-Gastrointestinal and Liver Physiology* 2017, **313**(3):G157-G165.
156. Luza SC, Speisky HC: **Liver copper storage and transport during development: implications for cytotoxicity**. *Am J Clin Nutr* 1996, **63**(5):812.
157. Blaner WS, Li Y, Brun PJ, Yuen JJ, Lee SA, Clugston RD: **Vitamin A Absorption, Storage and Mobilization**. *Subcell Biochem* 2016, **81**:95-125.
158. Spinella R, Sawhney R, Jalan R: **Albumin in chronic liver disease: structure, functions and therapeutic implications**. *Hepatol Int* 2016, **10**(1):124-132.
159. Lu H, Cassis LA, Kooi CWV, Daugherty A: **Structure and functions of angiotensinogen**. *Hypertens Res* 2016, **39**(7):492-500.
160. Winter WE, Flax SD, Harris NS: **Coagulation Testing in the Core Laboratory**. *Lab Med* 2017, **48**(4):295-313.
161. Vignesh P, Rawat A, Sharma M, Singh S: **Complement in autoimmune diseases**. *Clin Chim Acta* 2017, **465**:123-130.
162. Andersen CBF, Stødkilde K, Sæderup KL, Kuhlee A, Raunser S, Graversen JH, Moestrup SK: **Haptoglobin**. *Antioxid Redox Signal* 2017, **26**(14):814-831.
163. Hellman NE, Gitlin JD: **Ceruloplasmin metabolism and function**. *Annu Rev Nutr* 2002, **22**:439-458.
164. Gomme PT, McCann KB, Bertolini J: **Transferrin: structure, function and potential therapeutic actions**. *Drug Discov Today* 2005, **10**(4):267-273.
165. Thuener J: **Hepatitis A and B Infections**. *Prim Care* 2017, **44**(4):621-629.
166. Mitchell AE, Colvin HM: **Hepatitis and liver cancer: a national strategy for prevention and control of hepatitis B and C**: National Academies Press; 2010.
167. Aenishänslin HW, Stalder GA, Bianchi L, Gudat F, Carmann H: **[Hepatitis in drug addicts (follow-up study with biopsies) (author's transl)]**. *Dtsch Med Wochenschr* 1975, **100**(16):857-865.
168. Hosseini N, Shor J, Szabo G: **Alcoholic Hepatitis: A Review**. *Alcohol Alcohol* 2019, **54**(4):408-416.
169. Parker R, Kim SJ, Im GY, Nahas J, Dhesi B, Vergis N, Sinha A, Ghezzi A, Rink MR, McCune A *et al*: **Obesity in acute alcoholic hepatitis increases morbidity and mortality**. *EBioMedicine* 2019, **45**:511-518.
170. Barnett R: **Liver cirrhosis**. *Lancet* 2018, **392**(10144):31659-31653.

171. Hartke J, Johnson M, Ghabril M: **The diagnosis and treatment of hepatocellular carcinoma**. *Semin Diagn Pathol* 2017, **34**(2):153-159.
172. Razumilava N, Gores GJ: **Cholangiocarcinoma**. *Lancet* 2014, **383**(9935):2168-2179.
173. Young RJ, Brown NJ, Reed MW, Hughes D, Woll PJ: **Angiosarcoma**. *Lancet Oncol* 2010, **11**(10):983-991.
174. Sharma D, Subbarao G, Saxena R: **Hepatoblastoma**. *Semin Diagn Pathol* 2017, **34**(2):192-200.
175. Wong Y-H, Wu C-C, Lin C-L, Chen T-S, Chang T-H, Chen B-S: **Applying NGS data to find evolutionary network biomarkers from the early and late stages of hepatocellular carcinoma**. *BioMed Research International* 2015, **2015**.
176. Marengo A, Rosso C, Bugianesi E: **Liver Cancer: Connections with Obesity, Fatty Liver, and Cirrhosis**. *Annu Rev Med* 2016, **67**:103-117.
177. Maucort-Boulch D, de Martel C, Franceschi S, Plummer M: **Fraction and incidence of liver cancer attributable to hepatitis B and C viruses worldwide**. *Int J Cancer* 2018, **142**(12):2471-2477.
178. Hou W, Sanyal AJ: **Ascites: diagnosis and management**. *Med Clin North Am* 2009, **93**(4):801-817.
179. Portincasa P, Di Ciaula A, de Bari O, Garruti G, Palmieri VO, Wang DQ: **Management of gallstones and its related complications**. *Expert Rev Gastroenterol Hepatol* 2016, **10**(1):93-112.
180. Voloshina NB, Osipenko MF, Litvinova NV, Voloshin AN: **Hemochromatosis - modern condition of the problem**. *Ter Arkh* 2018, **90**(3):107-112.
181. Dyson JK, Beuers U, Jones DEJ, Lohse AW, Hudson M: **Primary sclerosing cholangitis**. *Lancet* 2018, **391**(10139):2547-2559.
182. Lindor KD, Gershwin ME, Poupon R, Kaplan M, Bergasa NV, Heathcote EJ: **Primary biliary cirrhosis**. *Hepatology* 2009, **50**(1):291-308.
183. Squires JE, McKiernan P, Squires RH: **Acute Liver Failure: An Update**. *Clin Liver Dis* 2018, **22**(4):773-805.

Chapter 2

Comparing *in vitro* human liver models to *in vivo* human liver using RNA-Seq

Rajinder Gupta[1], Yannick Schrooders[1], Duncan Hauser[1], Marcel van Herwijnen[1], Wiebke Albrecht[2], Bas ter Braak[3], Tim Brecklinghaus[2], Jose V. Castell[4], Leroy Elenschneider[5], Sylvia Escher[5], Patrick Guye[6], Jan G. Hengstler[2], Ahmed Ghallab[2,7], Tanja Hansen[5], Marcel Leist[8], Richard Maclennan[9], Wolfgang Moritz[6], Laia Tolosa[10], Tine Tricot[11], Catherine Verfaillie[11], Paul Walker[9], Bob van de Water[3], Jos Kleinjans[1], Florian Caiment[1, *]

1. Department of Toxicogenomics, School of Oncology and Developmental Biology (GROW), Maastricht University, Maastricht, The Netherlands
2. Leibniz Research Centre for Working Environment and Human Factors at the Technical University of Dortmund (IfADo) Dortmund Germany
3. Division of Drug Discovery and Safety, Leiden Academic Centre for Drug Research, Leiden University, PO Box 9503, 2300 RA, Leiden, The Netherlands
4. Experimental Hepatology Unit Instituto de Investigación Sanitaria La Fe Valencia Spain
5. Fraunhofer Institute for Toxicology and Experimental Medicine Preclinical Pharmacology and In-Vitro Toxicology Nikolai-Fuchs-Straße 1, 30625 Hannover, Germany
6. InSphero AG, Wagistrasse 27, 8952, Schlieren, Switzerland
7. Department of Forensic Medicine and Toxicology, Faculty of Veterinary Medicine, South Valley University, Qena 83523, Egypt.
8. In vitro Toxicology and Biomedicine, Dept inaugurated by the Doerenkamp-Zbinden foundation, University of Konstanz, Konstanz, Germany
9. Cyprotex Discovery, No 24 Mereside, Alderley Park, Cheshire, SK10 4TG, UK
10. Unidad Hepatología Experimental, Instituto de Investigación Sanitaria La Fe Valencia, Spain
11. Stem cell Institute, Department of development and regeneration, KU Leuven, Herestraat 49, 3000 Leuven, Belgium

Published: *Archives of Toxicology, 2020*

DOI: https://doi.org/10.1007/s00204-020-02937-6

Abstract

The liver plays an important role in xenobiotic metabolism and represents a primary target for toxic substances. Many different *in vitro* cell models have been developed in the past decades. In this study, we used RNA-sequencing (RNA-Seq) to analyze the following human *in vitro* liver cell models in comparison to human liver tissue: cancer derived cell lines (HepG2, HepaRG 3D), induced pluripotent stem cell derived hepatocyte-like cells (iPSC-HLCs), cancerous human liver derived assays (hPCLiS, human precision cut liver slices), non-cancerous human liver derived assays (PHH, primary human hepatocytes) and 3D liver microtissues. First, using CellNet, we analyzed whether these liver *in vitro* cell models were indeed classified as liver, based on their baseline expression profile and gene regulatory networks (GRN). More comprehensive analyses using non-differentially expressed genes (non-DEGs) and differential transcript usage (DTU) were applied to assess the coverage for important liver pathways. Through different analyses, we noticed that 3D liver microtissues exhibited a high similarity with *in vivo* liver, in terms of CellNet (C/T score: 0.98), non-DEGs (10363) and pathway coverage (highest for 19 out of 20 liver specific pathways shown) at the beginning of the incubation period (0h) followed by a decrease during long-term incubation for 168 and 336h. PHH also showed a high degree of similarity with human liver tissue and allowed stable conditions for a short-term cultivation period of 24h. Using the same metrics, HepG2 cells illustrated the lowest similarity (C/T: 0.51, non-DEGs: 5623, and pathways coverage: least for 7 out of 20) with human liver tissue. The HepG2 are widely used in hepatotoxicity studies, however, due to their lower similarity, they should be used with caution. HepaRG models, iPSC-HLCs, and hPCLiS ranged clearly behind microtissues and PHH but showed higher similarity to human liver tissue than HepG2 cells. In conclusion, this study offers a resource of RNA-Seq data of several biological replicates of human liver cell models *in vitro* compared to human liver tissue.

Introduction

The liver plays a central role in metabolizing exogenous substances. After oral uptake xenobiotics pass through the digestive tract and enter the liver via the portal vein, where metabolism by phase I and II enzymes take place [1]. Xenobiotics or their metabolites may damage the liver with fatal consequences for the individual [2]. Therefore, it is important to identify compounds that cause hepatotoxic effects to avoid exposure to humans.

While the use of animal models has proven to be of great importance in biological research [3-6], it remains challenging to translate the results to humans. Many drugs that showed great promise in animal testing failed safety assessment in clinical trials, e.g., emicizumab, zydelig, JCAR014, JCAR015, and Ad-RTS-hIL-12. To overcome these limitations human cell models have emerged as a viable alternative for efficacy, safety, and toxicity testing [7, 8]. These *in vitro* models do not just eliminate the species-specific variations but also have other advantages such as the requirement of only small amounts of the substance, relatively short testing periods, and the technically easy possibility to study mechanisms of toxicity, enzyme kinetics, and concentration-response relationships [7, 9]. Limitations using *in vitro* cell models are that differences between cells *in vitro* and *in vivo* may exist; moreover, relatively complex techniques are required to extrapolate from test compound concentrations in the culture medium *in vitro* to blood concentrations or doses *in vivo* [10, 11].

Several human liver cell models have been developed with an aim to resemble the *in vivo* situation as closely as possible[12]. HepaRG cells may be used for xenobiotic metabolism, toxicity studies, cytochrome P450 induction studies, and for analyzing genotoxic compounds [13, 14]. Primary human hepatocytes are still considered to represent a gold standard for hepatic biotransformation studies [15, 16], whereas HepG2 cells have been reported to represent a useful tool to study the regulation of drug-metabolizing enzymes [17]. In a review of different *in vitro* liver cell models, the advantages and disadvantages of the *in vitro* liver cell models have also been discussed [18]. Though informative, these studies only give a superficial comparison as they are based on selected processes and components, whereas next-generation sequencing (NGS) technologies can be used to obtain an unbiased, holistic view.

The evolution of NGS over the years has revolutionized genomics and transcriptomics research [19] making it affordable, fast, and precise. With NGS

based RNA sequencing (RNA-Seq), it has now become possible to both identify and quantify RNA transcripts [20], even in the absence of any prior genomic knowledge [19, 21]. Quantification of the transcript level, known as gene expression, can be analyzed in many different ways [7, 22-24] depending on the type of biological questions that need to be addressed. RNA-Seq provides the exhaustive expression profile of all genes expressed in the cell and is not limited to a set of genes widely studied.

In this study, we compared healthy human liver tissue, further referred to as "*in vivo* liver" with *in vitro* liver cell lines often used in toxicology studies. For the bioinformatics analysis, we used CellNet [25] which is a network biology-based computational platform to assess RNA-Seq expression data. In CellNet, consensus expression profiles of specific cells or tissue types were generated. For the ease of use, the authors have created transcriptome indices and annotation files of some cells/tissues by congregating publicly available RNA-Seq data for humans. We used these human indices and annotation files for comparing the liver *in vitro* cell models. Comparing the consensus expression data with the test cell models objectifies their similarity with different cells/tissues. CellNet also creates gene regulatory networks (GRN) that are derived from the expression profile. GRN is a network of genes that interact with each other to control specific cell functions [26]. GRNs can also be used to analyze similarities as they are specific for development, differentiation, and response to environmental cues [22].

In order to study each component (genes and/or transcripts) individually with equal weight, we also analyzed non-differentially expressed genes (non-DEGs). Usually, differentially expressed genes (DEGs) between samples are analyzed to describe the differences between cell types, exposures, time-points, or other influences [22]. Here, also the non-DEGs were analyzed to focus on the similarities between *in vitro* liver cell models and *in vivo* liver. The higher the number of non-DEGs the higher the similarity an *in vitro* model and the liver.

Gene expression levels from RNA-Seq data are usually obtained by summing the reads attributed to all transcript (or isoforms) variants for each given gene so that the change in the amount of expression of individual isoforms is not apparent. A previous consensus has been that the majority of genes are regulated through their mRNA levels but NGS has shown that also the selection of individual spliced variants may change while the sum of all isoforms remains unchanged. Moreover, the ENCODE project revealed through NGS that close to 95% of human multi-exon

genes undergo alternative splicing [27] to form the gene transcripts. Gene transcripts are mRNAs that have different transcription start sites (TSSs), protein coding DNA sequences (CDSs), and/or untranslated regions (UTRs) but all are expressed from the same locus. Ensembl [28] provides an extensive list of transcript types broadly categorized as protein coding, nonsense mediated decay, non-stop decay, and long as well as small non-coding RNA (http://www.ensembl.org/info/genome/genebuild/biotypes.html). By RNA-Seq, quantified information of different transcripts of a gene can be obtained [19] and changes in the fraction of each transcript, known as differential transcript usage (DTU), can be studied to provide insights. These differences in the ratio of the expression of the transcripts can potentially alter the gene function and the mRNA regulation, stability, and localization [29, 30].

To our knowledge, previous studies have compared genome-wide expression only of individual cell models, such as e.g. PHH and iPSC derived human hepatocyte-like cells [16, 31]; however, a systematic comparison of the most frequently applied *in vitro* liver cell models to human liver tissue has not yet been performed. Here, we studied different *in vitro* liver cell models at baseline conditions, i.e. without any compound exposure: liver models derived from cancer cells (HepG2, HepaRG 3D), iPSC (induced pluripotent stem cells) derived hepatocyte like cells, cancerous human liver derived (hPCLiS, human precision cut liver slices), and non-cancerous human liver derived cultivated primary human hepatocytes (PHH) and 3D liver microtissues. These cell models were compared to healthy *in vivo* liver assessed by NGS data.

Materials and methods

An overview of the analyzed samples, human liver tissue specimens *in vivo* and *in vitro* liver cell models is given in Table 1, detailed information on samples and protocols is available in Suppl. methods 1a-g, and details on samples selected after each filtration step are provided in Suppl. Table 1. For PHH, hPCLiS, and 3D liver microtissues expression data for multiple time points are available.

***Table 1**: Overview of in vitro liver cell models used in this study.*

Cell line	Cultivation period (in hours)	No of replicates (biological/technical)			Protocol and other details
		Total	Sequencing depth filtration	After bootstrapping (replicates most similar to *in vivo* healthy liver)	
In vivo liver	NA	27	24[#]	24	Suppl. methods 1(a)
PHH	0*	6	6	3	Suppl. methods 1(b)
	24	6	6	3	
iPSC-HLC	480	6**	3	3	Suppl. methods 1(c)
3D liver microtissues	0*	3	2	2	Suppl. methods 1(d)
	168	3	3	3	
	336	3	3	3	
HepG2	0*	7	7	3	Suppl. methods 1(e)
HepaRG 3D	0*	4	3	3	Suppl. methods 1(f)
hPCLiS	0*	4	4	3	Suppl. methods 1(g)
	24	4	4	3	

*Timepoint 0 hours is time post-seeding

**Donor 1 (SBAD2) → 4 replicates; Donor 2 (SBAD3) → 2 replicates

[#]3 samples from infants or children were removed

RNA sequencing

All samples from the *in vitro* liver cell models were analyzed by a standardized working pipeline that included the immediate transfer of cells and tissues into TRIzol™ after the cultivation periods as indicated in Table 1. RNA was extracted from these cell samples with a Qiagen miRNeasy Mini Kit (Cat # 217004). Additionally, DNase digest was performed with a Qiagen RNase-Free DNase Set (Cat # 79254) to remove unwanted DNA. RNA quantity and quality were assessed by Qubit™ RNA HS Assay Kit (Cat # Q32855) and Agilent RNA 6000 Nano Kit (Cat #5067-1511) respectively and prepared for sequencing with the Lexogen SENSE mRNA-Seq Library Prep Kit V2 (Cat # 001.96). After library preparation was completed, the

quality of the libraries was checked on an Agilent 2200 TapeStation using an Agilent High Sensitivity D5000 ScreenTape (Cat # 5067-5592) and library concentration was determined by Qubit™ dsDNA HS Assay Kit (Cat # Q32854) before proceeding to sequencing. While the healthy liver tissue samples were sequenced (paired-end, 150bp) on an Illumina NovaSeq 6000® using a single S2 flowcell, the *in vitro* cell models were sequenced (paired-end, 100bp) on Illumina HiSeq2000®.

Data pre-processing

The quality of the RNA-Seq raw data (fastq files) was analyzed using the Fastqc (version 0.10.1) [32] and after considering the quality of the sequences, tails of the sequences were trimmed of the bad quality of the sequences (twelve nucleotides) using Trimmomatic (version 0.33) [33]. The sequences were mapped onto the Ensembl [28] human genome (version 84) using Bowtie2, (version 2.2.6) [32], and gene and isoform (transcript) expression were calculated using RSEM (version 1.2.28) [34]. Using the sorted genome bam files from RSEM, annotation of the mapped reads was assessed by applying ALFA [35].

The gene read counts, isoform read counts and isoform percentage from all *in vivo* and *in vitro* samples were taken. Gene read counts were used for finding the non-DEGs, then isoform count and percentage were used to analyze DTU. Calculation of non-DEGs and DTU is done for each cell model and all timepoints, individually.

CellNet analyses

The fastq files were used for CellNet analysis. All the subsequent steps were performed locally as explained in the CellNet protocol paper [36]. We used the 'Human Jun_20_2017' cnProc from the CellNet for analyses. Two types of analyses were performed: comparing the consensus expression profile and GRN status. The consensus expression profile per cell or tissue type is generated from publicly available RNA-Seq data and classification scores for the test samples are obtained. GRN created from the consensus expression profile give the GRN status when samples are compared against them. These are calculated by first computing the raw GRN status as the mean z-score of all genes in a C/T (cell/tissue) GRN, weighted by their importance to the associated C/T classifier. The raw GRN status is then

normalized to the mean raw GRN status of the training data samples of the given C/T [37].

Bootstrapping

A different number of replicates (technical or biological) were present for all cell models. To eliminate this possible source of bias, we selected three replicates from each cell model which presented the highest similarity with the healthy *in vivo* liver (Table 1, Suppl. Table 1) based on the number of non-DEGs. In the case of 3D liver microtissues, only two replicates were taken instead of three because the third replicate had very low coverage and was discarded at the sequencing depth filtration. These selected replicates were then used to calculate non-DEGs, DEGs, DTU, and other further analyses.

Non-DEGs

The data were normalized by defining *in vivo* liver as one dataset and each *in vitro* liver model for each time point as an individual dataset (best three replicates were taken as explained above). Then each *in vitro* dataset was compared individually to the *in vivo* dataset for calculating the non-DEGs using the 'DESeq' function from DESeq2 R-package [38]. The list of non-DEGs is obtained by filtering the results for q-value (padj) > 0.05 and basemean > 10. These non-DEGs were mapped onto KEGG pathways [39] using Pathview [40], and in-house developed scripts were used to calculate the pathway coverage.

Differential transcript usage (DTU)

The change in the proportion of the transcripts expressed for a gene represents differential transcript usage. Isoform counts and percentages were calculated using RSEM. The isoform counts were normalized using DESeq2 as explained for gene reads for the selected replicates for each cell model. Considering the number of transcripts assessed, multiple filtering steps were applied to remove the low expressed transcripts (or noise), and transcripts expressed at a similar level from the control (*in vivo*) and test (*in vitro*) samples.

i. Low expression/noise:

Isoforms that were expressed less than one in a million reads in one dataset (test or control) were removed. These isoforms were removed because their expression level was not sufficient to be considered above the noise at this sequencing depth. This filtration step was performed on isoform counts.

ii. Similar expression:

Isoforms that differ less than equal to 10% between the average percentage of *in vivo* and *in vitro* samples were removed because we were interested in looking for the isoforms having sufficient differential usage. This filtration step was performed on isoform percentages.

iii. No expression in some samples:

The isoforms that were not detected in more than 20% of the samples in any one of the datasets (*in vivo* or *in vitro*) were discarded, as this would reduce the confidence in the samples that showed expression for those isoforms. If the number of samples for test or control were less than five, we imposed that the transcripts were detected in all samples. This filtration step was performed on isoform percentages.

Isoforms deleted from the counts dataset were removed from the percentages dataset and the ones deleted from the percentages were removed from the counts.

After the filtration steps, the genes left with only one isoform were removed from both datasets (counts and percentages). The variance was calculated between the test and control samples for all the remaining transcripts using ANOVA in R. Isoform percentages were used to find the variance because percentages are linearly distributed (contrary to the RNA-Seq read count). It was filtered on p-value < 0.01 as the calculation at the isoform level has a higher error rate.

The highest expressed isoform was identified (highest percentage) in the control samples for each gene and was compared with its expression profile in test samples. The genes with different expression profiles (DTU) were removed from the non-DEGs and were named as non-DEGs^{DTU-}. The list of non-DEGs^{DTU-} was mapped onto the KEGG pathways and pathway coverage was recalculated.

Results

RNA from totally of 27 liver tissue specimens from donors without liver diseases, further named “healthy in vivo liver” and 46 samples from cultivated hepatocytes, cell lines, liver slices or iPSC derived hepatocyte-like cells, so-called *in vitro* cell models, were sequenced on the Illumina NovaSeq (PE, 150bp) and the Illumina HiSeq 2000 (PE, 100bp), respectively. After removing the samples from children or infants for healthy *in vivo* liver specimens and filtering for sequencing depth 24 *in vivo* cell models, 38 *in vitro* samples remained for further analysis (Suppl. Table 1). Since healthy human liver represents a very valuable resource, we generated sequences at very high depth ($1.63*10^8$) for community usage. However, to avoid coverage bias in our analysis with the *in vitro* samples (sequenced at a depth of $33.65*10^6$), only the first 30 million reads of the fastq files obtained from the *in vivo* samples were used. We compared the full coverage and part of the data and found that the whole sample and sub-selection had similar distribution (Suppl. Fig. 1).

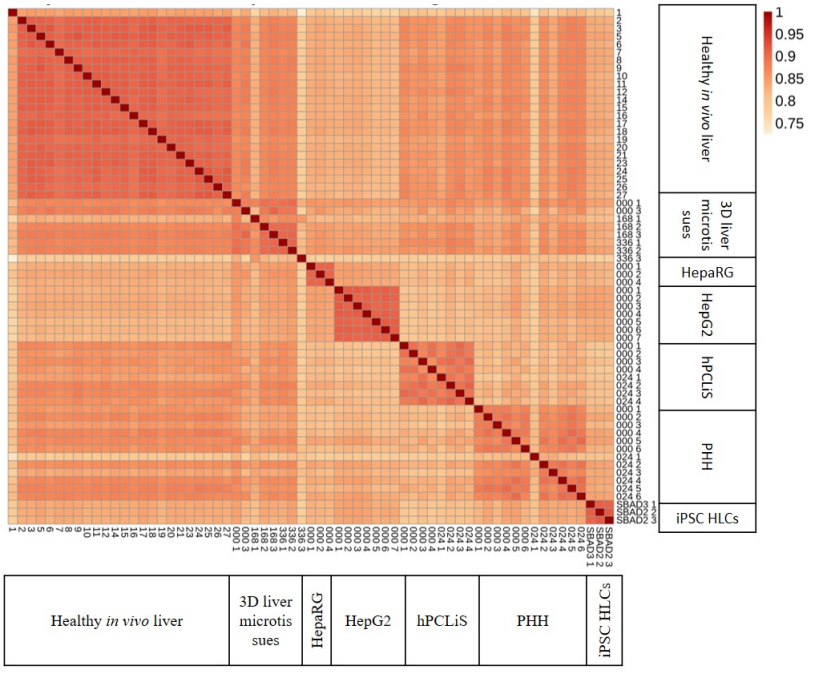


***Figure 1**: Spearman’s correlation plot. The Spearman’s correlation plot for normalized read counts of all in vivo and in vitro samples taken after first sample filtration. For healthy in vivo liver, the replicate numbers are given. For all in vitro cell models cultivation periods*

(000/024/168/336 h) and replicate numbers are indicated except iPSC HLC. In the case of iPSC HLC the donor id (SABD2/3) and replicate number are given. The color bar indicates the Spearman correlation coefficient of each pairwise correlation.

All samples taken after initial filtration passed the 'per base sequence quality' metric of the Fastqc. The annotation of reads was assessed using ALFA. Median protein coding reads were 47.02%, 52.3%, 45.52%, 40.9%, 40.87%, 35.15%, and 47.16% for healthy liver tissue, 3D liver microtissues, HepaRG, HepG2, hPCLiS, PHH, and iPSC-HLC, respectively (Suppl. Fig. 2a and 2b). The samples that had lower protein coding and 3'UTR reads showed an increase in the intergenic and 5'UTR reads. Overall, the samples had similar distributions across different regions. Furthermore, the global similarity of the cell models was evaluated using pairwise Spearman's correlation for normalized read counts (Fig. 1). The median (and standard deviation) of all pairwise correlation coefficients of the healthy liver tissue specimens with the samples from 3D liver microtissues, HepaRG, HepG2, hPCLiS, PHH, and iPSC-HLC, were 0.87 (0.033), 0.83 (0.008), 0.82 (0.013), 0.86 (0.014), 0.86 (0.014), and 0.83 (0.01), respectively. The variation coefficients of 3D liver microtissues, HepaRG, HepG2, hPCLiS, PHH, and iPSC-HLC, were 3.87, 0.99, 1.57, 1.65, 1.59, and 1.22, respectively. Inter-replicate variation was observed predominantly in 3D liver microtissues (166_1 and 336_3) and PHH (024_1) cell models. It should be considered that interindividual variability contributes to the cell models obtained from different donors (healthy liver tissue specimens, 3D liver microtissues, hPCLiS, PHH), in contrast to the cell line derived cell models (iPSC-HLC, HepaRG, HepG2).

CellNet cell/tissue classification scores of RNA-Seq expression profiles

Since the quality and global distributions of the samples were comparable, we next assessed their transcriptome expression to consensus profiles of different cells and tissues. Using CellNet on the expression data of our liver *in vivo* and *in vitro* samples, we calculated classification scores (Fig. 2 and Suppl. Table 2). CellNet classified all cell models as liver. We noticed that the iPSC-HLCs present the lowest CellNet classification score for the human liver and they still share some resemblance with embryonic stem cells (ESC). Among all the cell models, HepaRG 3D had the highest classification score for fibroblasts. The cancer cell models (HepG2 and HepaRG 3D) also exhibited low classification scores as compared to non-cancerous liver-derived

cell models, whereby the classification scores of HepaRG 3D were slightly higher compared to HepG2 cells but still much lower than the values for 3D liver microtissues, hPCLiS, and iPSC-HLC. The human liver derived models (3D liver microtissues, hPCLiS, and iPSC-HLC) did not show major differences among each other based on the CellNet classification score. Furthermore, at the level of GRNs (status score) (Suppl. Fig. 3) similar results were obtained as for the consensus expression comparison (classification score). CellNet results can be used to find the extent of similarity and dissimilarity for the cell models but other approaches should be used to identify the differences at gene and transcript level. In this study, we explored non-differentially expressed genes (non-DEGs) and differential transcript usage (DTU) to provide comprehensive comparisons between the cell models. However, to remove the bias caused by the different number of replicates from each cell model, we performed bootstrap analyses to guarantee that an identical number of replicates from each cell model was used (Fig. 3, Suppl. Table 1). The cell models had differing number of replicates, hence the number of combinations of replicates, taken three at a time, also varied across cell models. The number of DEGs for 3D liver microtissues were 6315 (0h), 9552 (168h), and 9478 (336h), for iPSC HLCs it was 12155 (480h), and 11790 for HepaRG 3D. For these cell models, only one combination of replicates per time point was obtained. For the remaining cell models, where the number of replicates were more than three after the initial filtration for quality, the average number of DEGs for PHH were ~9684 (0h) and ~9508 (24h), for hPCLiS the mean was ~12499 (0h) and ~12815 (24h), and for HepG2 it was 13070 (0h). From these, the best three replicates were selected based on the number of non-DEGs, except for 3D liver microtissues 0h because one of the replicates was discarded for low coverage.

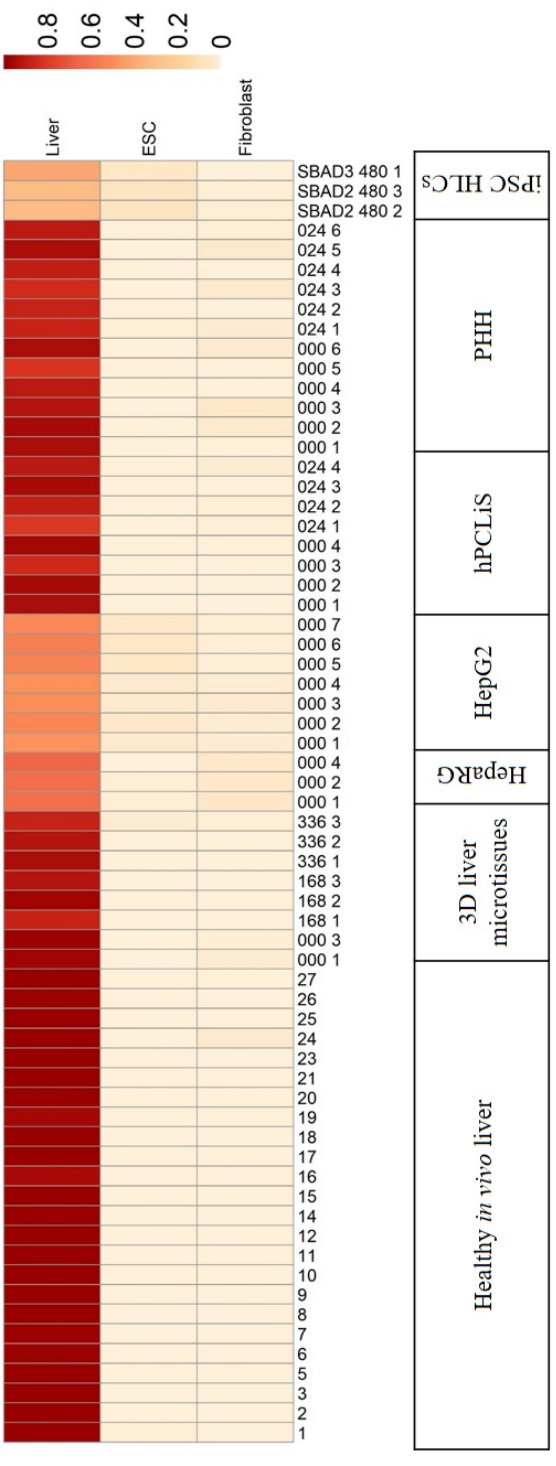
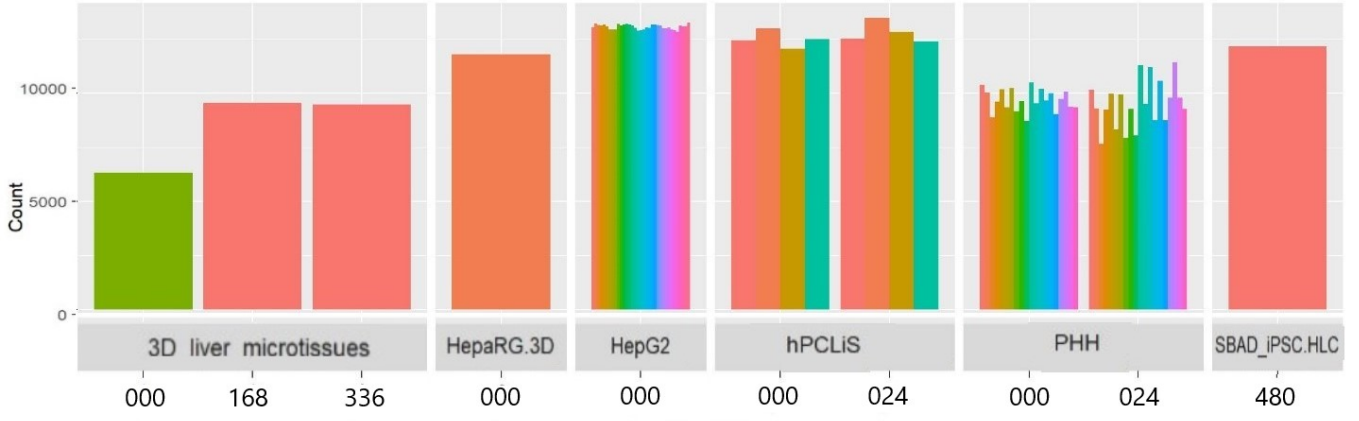


Figure 2 *: CellNet C/T classification score. The classification score for in vivo and in vitro samples compared to the liver, embryonic stem cell (ESC), and fibroblast data of CellNet, represented as a heat map. For healthy in vivo liver, the replicate numbers are indicated. For all in vitro cell models, time points (000/024/168/336 h) and replicate numbers are given except iPSC HLC. In the case of iPSC HLC, the samples are labeled as donor id (SABD2/3) and replicate number. The color bar represents the classification score as calculated by CellNet.*



Number of combinations for the replicates

	000	024	168	336	480
3D_liver_microtissues	1	-	1	1	-
HepaRG.3D	1	-	-	-	-
HepG2	35	-	-	-	-
hPCLiS	4	4	-	-	-
PHH	20	20	-	-	-
SBAD_iPSC.HLC	-	-	-	-	1

Figure 3*: Number of DEGs from combination of replicates during bootstrapping. There were variable number of replicates for the cell models. This may result in incomparable statistical analyses of the cell models. Therefore, to address this concern, a bootstrap strategy was applied to select the replicates that had the least number of DEGs when compared to in vivo liver. Different colors of the bar graph represent various combinations of the replicates.*

***In vivo* versus *in vitro*, using non-differentially expressed genes**

Normalized gene expression (mRNA profiles) of the *in vitro* cell models and *in vivo* liver were used to identify genes that are not differentially expressed (non-DEGs) (Fig. 4 and Suppl. Table 3) to characterize which liver-like features the individual cell models possess. The numbers of non-DEGs of 3D liver microtissues (0h) before cultivation was highest compared to the other cell models but dropped below the corresponding numbers of PHH after long-term cultivation for 168 and 336h (Fig. 3). Similar numbers of non-DEGs were obtained for PHH before and after short-term cultivation for 24h. The lowest numbers of non-DEGs were obtained for HepG2 and hPCLiS, while HepaRG and iPSC-HLC were intermediate. The highest overlap of non-DEGs was obtained between PHH before and after the cultivation period, illustrating that this system offers a relatively stable number of non-DEGs during short-term incubation for 24h (Fig. 3). Moreover, a relatively large overlap of non-DEGs was obtained for iPSC-HLC and PHH.

To understand the effect on the biological processes, we then mapped all non-DEGs onto KEGG pathways (Suppl. Table 4a). Pathway mapping data can be used to study the specific processes/pathways of interest for each cell model and provide a metric of the similarity between liver tissue and the individual *in vitro* systems. Pathway coverage was calculated for the 20 liver pathways [41] illustrated in Fig. 5 (Suppl. Table 5). Higher pathway coverage by non-DEGs implies higher similarity with the human liver.

3D liver microtissues (0h) before the incubation period showed the highest coverage for most pathways but after 168 and 336 h of incubation, the coverage systematically dropped. In general, HepaRG 3D and HepG2 demonstrated a much lower coverage with HepG2 having the least. Exceptions were the high DNA repair functions of both tumor cell lines, with a relatively high coverage seen for base excision repair for HepaRG (68%) and nucleotide excision repair for HepG2 (59%). PHH showed a relatively high pathway coverage for all pathways and only small differences before (0h) and after (24h) after the cultivation period. For the primary bile acid biosynthesis pathway, PHH showed an increase during the 24h cultivation period. For cytochrome P450 pathways, microtissues, PHH, and hPCLiS demonstrated a high coverage (metabolism of xenobiotics: hPCLiS 0h: 71%, 24h: 79%, PHH 0h: 79%, 24h: 71% and drug metabolism: hPCLiS 0h: 78%, 24h: 94%, PHH 0h: 89%, 24h: 89%), hence presenting their metabolizing capacities for exogenous substances. While the number of pathways for which hPCLiS exhibited a low

coverage, was six for 0h and eight for 24h, it also had the highest coverage for two pathways for 24h (both cytochrome p450 pathways). IPSC-HLC also showed a high coverage for DNA repair pathways with highest on base excision repair (68%) and nucleotide excision repair (66%). For other DNA repair pathways, iPSC-HLC also exhibited a higher coverage than HepaRG and HepG2 cell models.

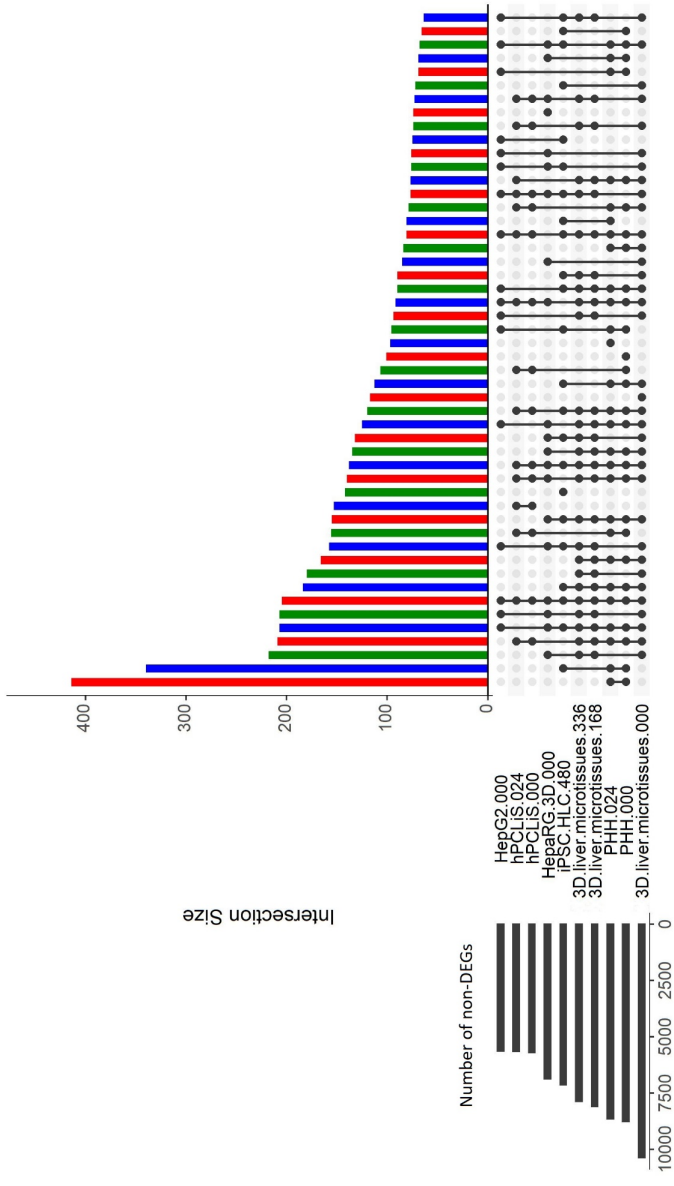


***Figure 4**: Overlap and number of non-DEGs. The number of non-DEGs for all cell models obtained after comparing against in vivo samples shown as horizontal bar plots on the left. The overlap between all cell models is shown as the main graph, top 50 overlaps are shown. For each cell model, the best three replicates were chosen as explained in the Bootstrapping section under Materials and methods. Different colors are used to enhance the readability of the graph.*

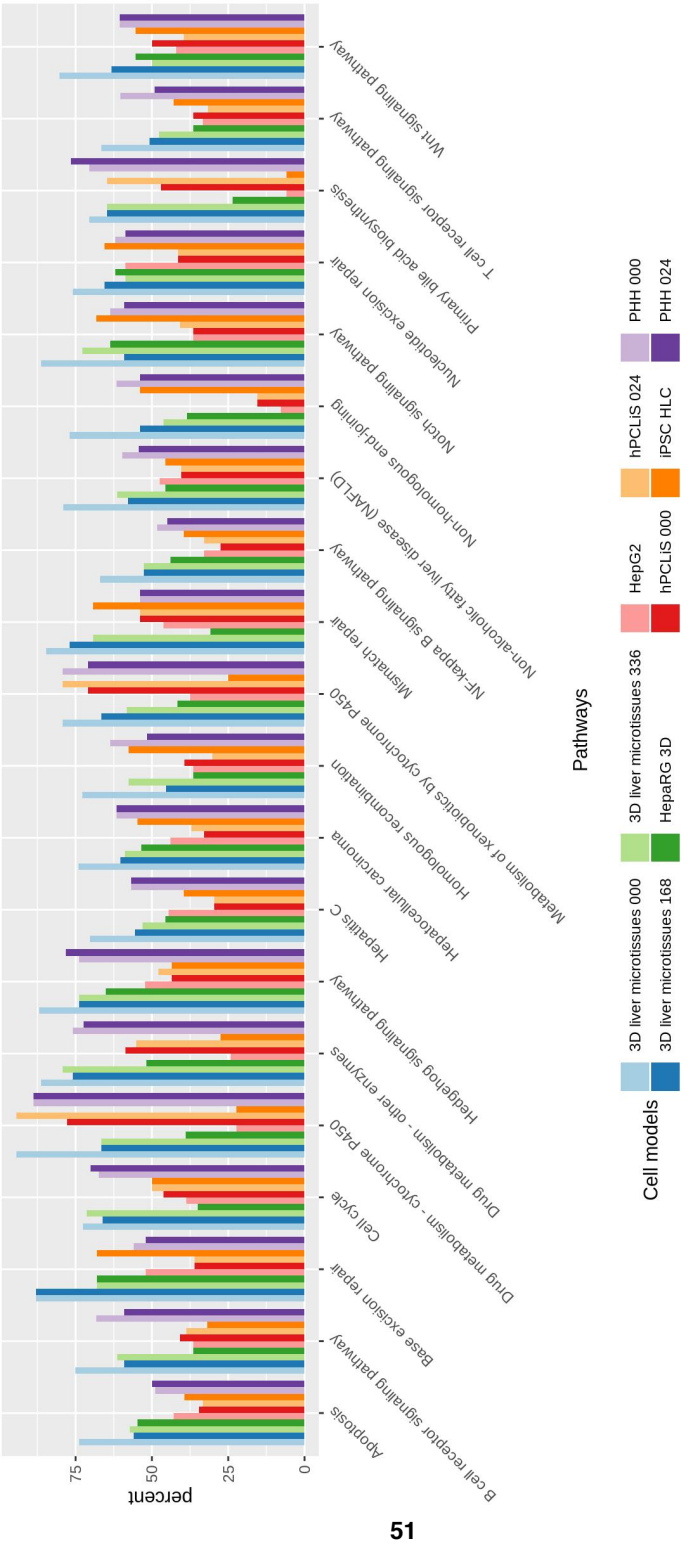
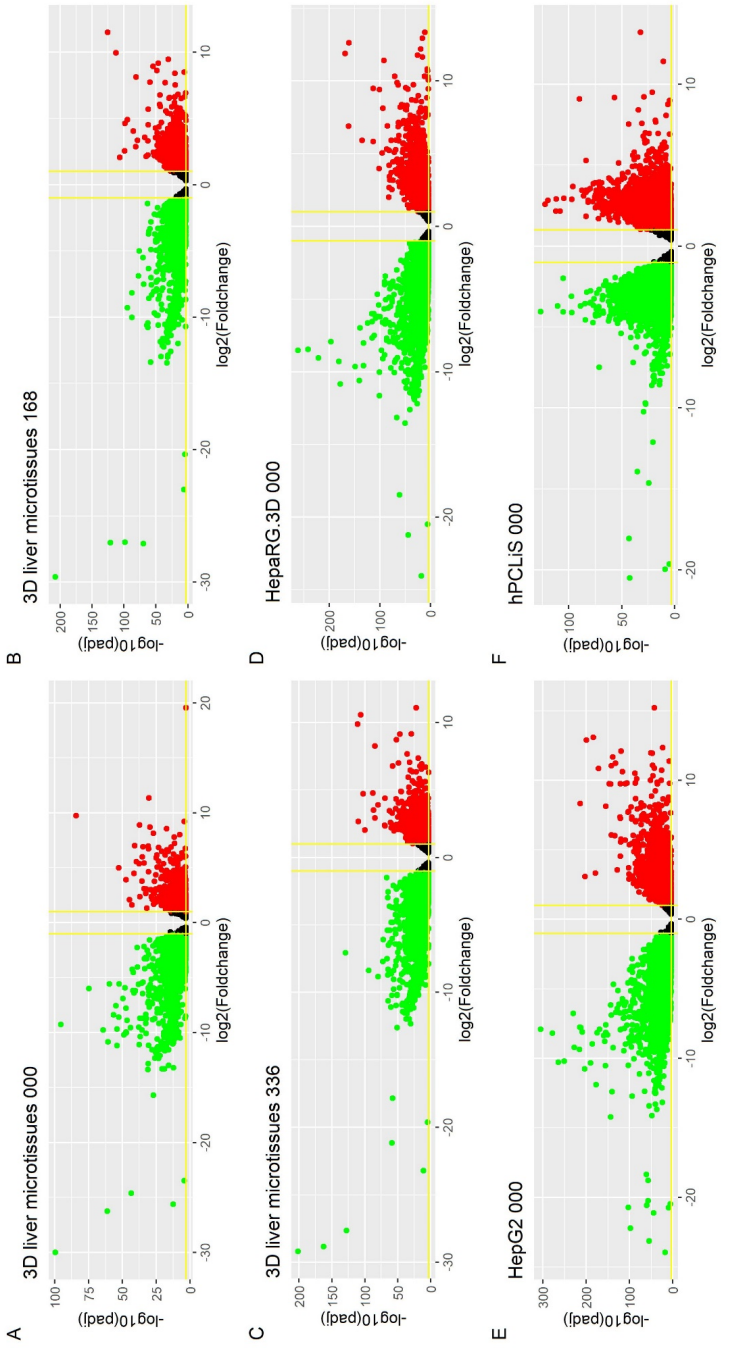


***Figure 5**: Pathway coverage of liver pathways by different cell models for the non-DEGs. The non-DEGs from all in vitro cell models were mapped onto important pathways in the liver for cell processes, regrowth and regeneration, cancer, viral infection, immune response, drug and xenobiotics metabolism, repair, and toxicity.*

***In vivo* versus *in vitro*, using differentially expressed genes**

While the non-DEGs illustrated the similarities between the *in vitro* cell models and the *in vivo* liver, we also compared the differentially expressed genes (DEGs, q-value (padj) < 0.05 & average counts > 10) to highlight the differences. The volcano plots demonstrated the extent of perturbation in the genes for all cell models (Figure 6A-J). The number of DEGs were the highest for HepG2 (9910) and lowest for 3D liver microtissues 000 (5169) (Figure 6K, Suppl. Fig. 4). The number of DEGs were also high, comparable to HepG2, for hPCLiS both time points (0h: 9837 and 24h: 9890). The complete list of DEGs from all cell models is provided in Suppl. Table 6. The overlap between the DEGs from all cell models in Suppl. Fig. 4 shows that the highest overlap was between all cell models except both time points from PHH. An enrichment analyses was performed for the DEGs using GOrilla [42] (Suppl. Table 7). While iPSC, HepaRG and HepG2 demonstrated the most perturbed GO functions, PHH had the least (Fig. 7). The highest overlap (19 GO functions) was between the iPSC and HepG2 cell models.


A
3D liver microtissues 000
B
3D liver microtissues 168
C
3D liver microtissues 336
D
HepaRG.3D 000
E
HepG2 000
F
hPCLiS 000
-log10(padj)
log2(Foldchange)

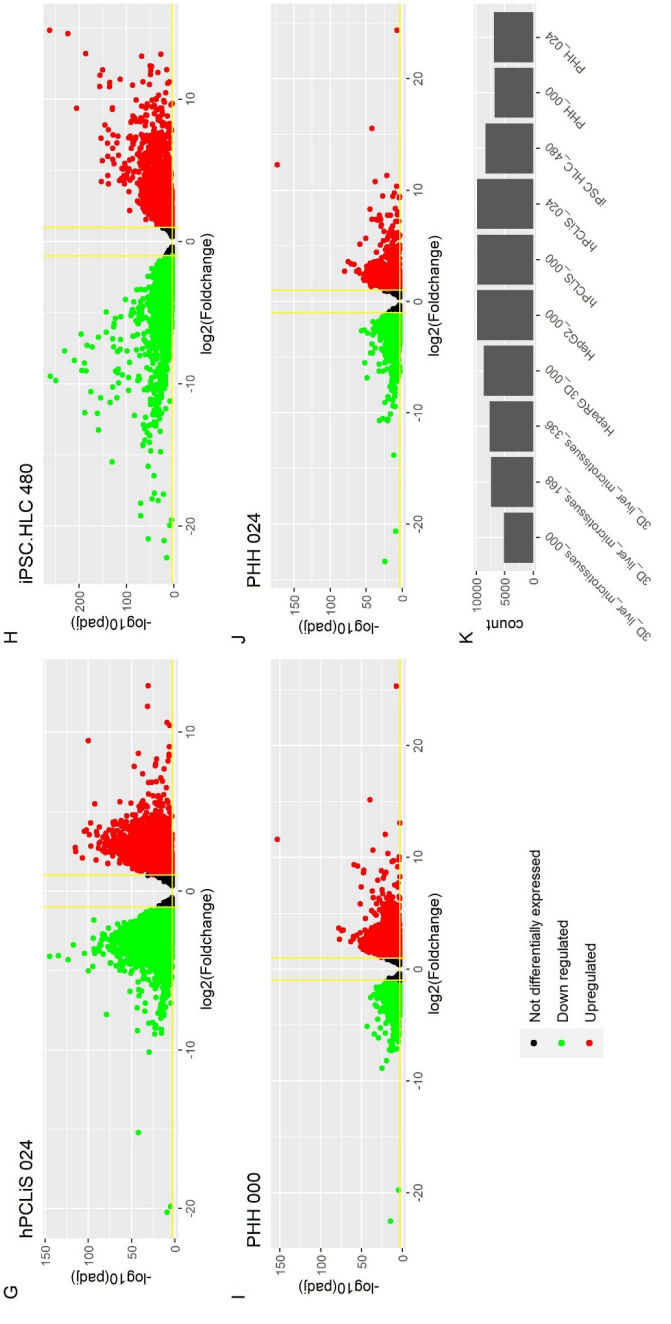


***Figure 6**: Volcano plots for DEGs. (A-J) The DEGs from various cell models when compared with healthy in vivo liver. The black dots represent not differentially expressed, green dots down regulated and red dots up-regulated genes. The x-axis is the log2foldchange of the gene expression between the healthy and in vitro cell models and y-axis is the p-adjusted (padj or q-value). The horizontal yellow line corresponds to –log10(0.05) where 0.05 is the threshold for padj and the vertical lines correspond to log2foldchange < -1 and > 1. (K) Number of DEGs from each comparison.*

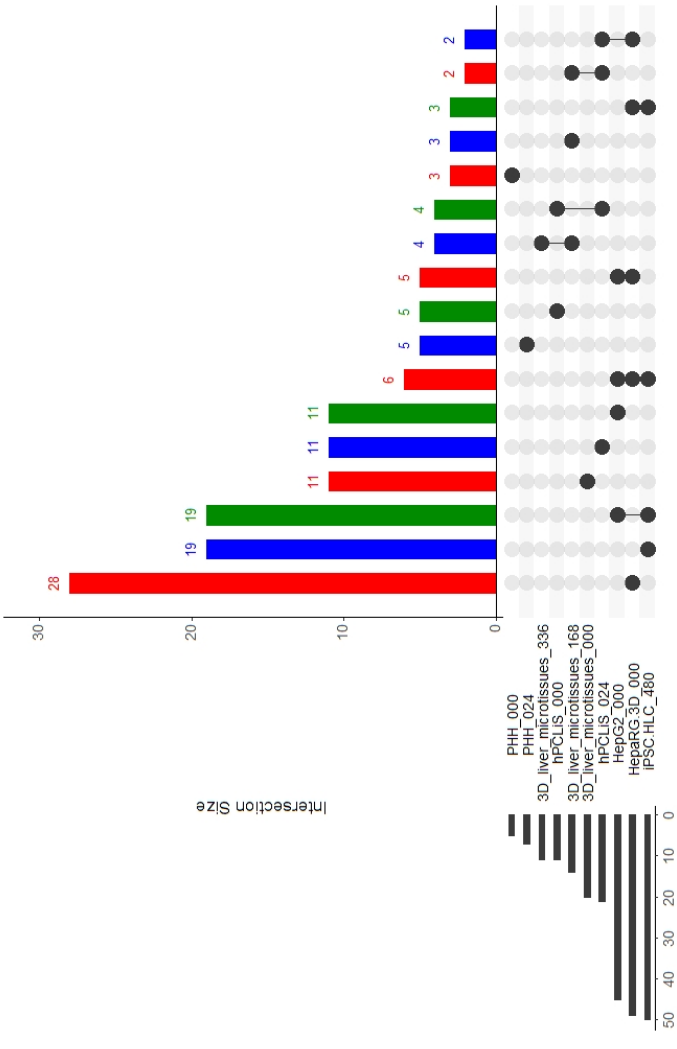


***Figure 7**: Overlap between the GO function for the DEGs. An enrichment analyses for the DEGs from all in vitro cell models was performed and the overlap for the resulting GO functions is presented. Different colors are used to enhance the readability of the graph.*

The DEGs were also mapped onto the pathways to check their coverage (Fig. 8). A higher coverage by DEGs means that the cell models share low similarity with healthy *in vivo* liver. It is important to mention here that the pathway coverage for the DEGs is not the inverse of the pathway coverage of non-DEGs, this is because different genes can make similar proteins. The pathways are proteins interacting with each other and due to ambiguity in protein-gene relationships, pathway mapping tools, frequently map more than one gene to a protein. The pathway coverage of the DEGs illustrated an opposite mapping trend than non-DEGs (Fig. 5) and correctly so. Overall HepG2, HepaRG, and hPCLiS illustrated the highest

coverage whereas PHH and 3D liver microtissues showed lowest coverage and iPSC HLCs had high for some and low coverage for other pathways. Pathway mappings on DEGs from all cell models for all human pathways are provided in Suppl. Table 4b.

The changes in the expression of genes, differentially expressed genes, can be linked to the fluctuations in the expression of different transcription factors (TFs). The Network influence score (NIS), defined by the expression of downstream regulated genes, for the transcription factors from all the cell models was calculated using CellNet (Suppl. Fig. 5a-g). These differences were calculated with respect to the cell/tissue profiles of the CellNet. The results show that the transcription factor ATF5 had the highest perturbation for all cell models except the PHH where it was shown to be the least perturbed. In the case of PHH, NR1H4 was the most affected factor. Moreover, in the case of PHH, ATF5 exhibited perturbation in the opposite direction than all other cell models. A similar analysis using the microarray data for freshly extracted hepatocytes, PHH and hiPSC using the microarray data illustrated a different list of TFs being affected. However, different types of data used for the two studies (microarray and transcriptomics) might be the reason for this difference.

Furthermore, we also investigated how the cell models behaved over the incubation period. Three cell models, namely, PHH, hPCLiS, and 3D liver microtissues were incubated for different time durations. We computed the DEGs for each cell model over different time points (Fig. 9). PHH and hPCLiS show only very small variation in their expression profile over time, with a single gene differently expressed for PHH (0h vs 24h) and two DEGs for hPCLiS (0h vs 24h). 3D liver microtissues show a more important effect of time, with 684 DEGs between 0h and 168h, 223 between 0h and 336h and 8 for 168h vs 336h. While 3D liver microtissues illustrated comparatively higher number of DEGs, it should also be acknowledged that the incubation period for 3D liver microtissues was much longer than PHH and hPCLiS. PHH and hPCLiS that had same incubation period showed only a few genes perturbed over time.

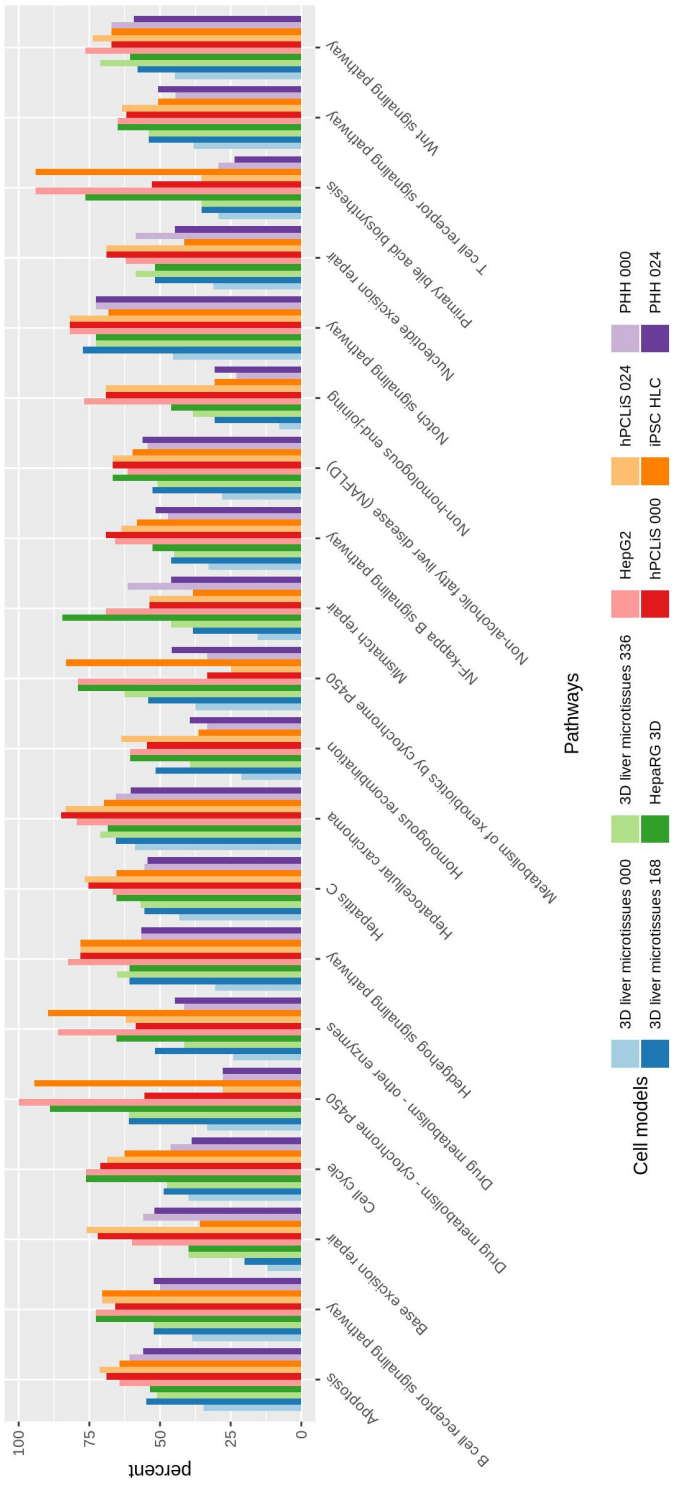


***Figure 8**: Pathway coverage of liver pathways by different cell models for the DEGs. The DEGs from all in vitro cell models were mapped onto important pathways in the liver for cell processes, regrowth and regeneration, cancer, viral infection, immune response, drug and xenobiotics metabolism, repair, and toxicity.*

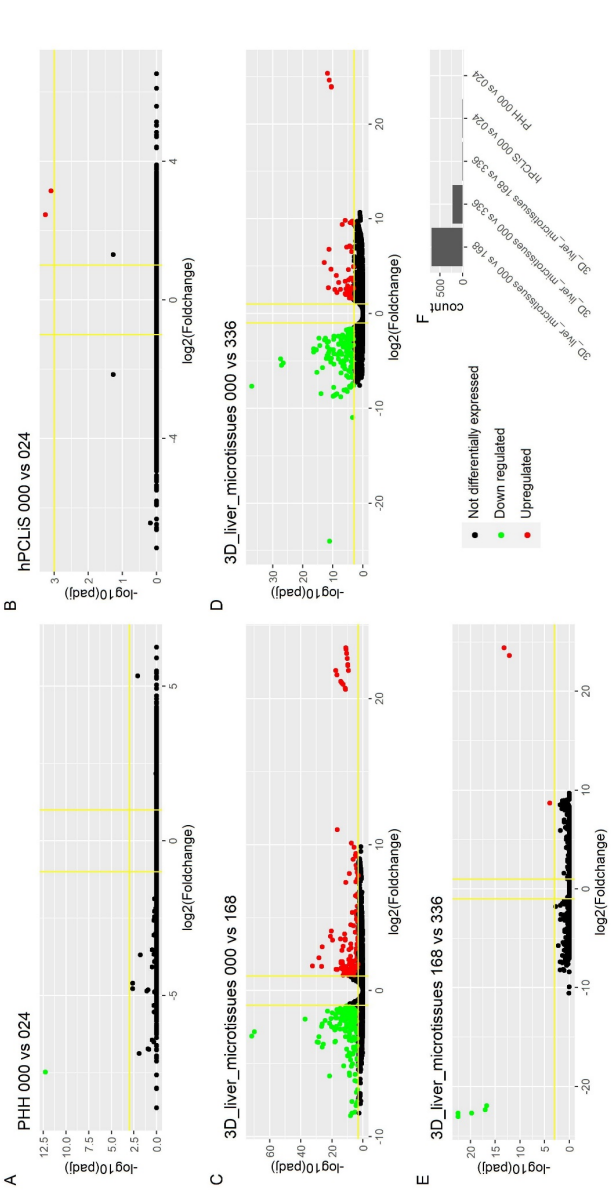


***Figure 9**: Volcano plots for differentially expressed genes (DEGs) from comparison of incubation periods for the cell models. RNA-Seq data was available for different time points for PHH (0 and 24h), hPCLiS (0 and 24h), and 3D liver microtissues (0, 168, and 336h). DEGs were computed between different incubation times for each cell model. (A-E) Volcano plots for the DEGs computed for different comparisons. The black dots represent not differentially expressed, green dots down regulated and red dots up-regulated genes. The x-axis is the log2foldchange of the gene expression between the healthy and in vitro cell models and y-axis is the p-adjusted (padj or q-value). The horizontal yellow line corresponds to –log10(0.05) where 0.05 is the threshold for padj and the vertical lines correspond to log2foldchange < -1 and > 1. (F) Number of DEGs from each comparison.*

***In vivo* versus *in vitro*, using differential transcript usage**

In the previous analyses, RNA-Seq data have been analyzed to identify differences between *in vivo* and *in vitro* cell models for the total gene expression generated by all isoforms of a gene. If the proportion of expression changes between different isoforms of a gene, total gene expression may remain constant. However, the different transcript usage (DTU) may nevertheless be relevant, because DTU may generate functionally different gene products. Differences in transcript expression (DTU) may be caused by alternative splicing, preference for one transcription start site over the other, spatial availability of transcription factors, and other elements. In standard RNA-Seq analysis, gene expression is assessed by summing the expression of all the transcripts for a given gene, and then it conceals the genes regulated at the splicing level. Genes with significant differential usage (p-value < 0.01) at transcript level were then removed from the list of non-DEGs, giving the non-DEG^{DTU-} (Suppl. Table 8). An exhaustive list of DTU for all the cell models can also be found in the supplementary (Suppl. Table 9). The gene for which transcripts had differential usage (DTU) should be removed from the non-DEGs to fine-tune the analyses. This was illustrated for the examples of four genes that were non - differentially expressed but display DTU for the *in vitro* cell models (Fig. 10). The highest expressed protein coding transcript of *POLR2F* (DNA-directed RNA polymerases I, II, and III subunit *RPABC2*) was mostly replaced by other protein coding transcripts. For *GOLGA8B* (Golgin subfamily A member 8B) and *ARHGAP21* (Rho GTPase-activating protein 21), it was predominantly replaced by non-coding transcripts. *HSPA8* (Heat shock cognate 71 kDa protein) exhibited a different pattern, where the highest expressed protein coding transcript was replaced by other protein coding and non-coding transcripts. The highest expressed protein coding transcript in the case of *POLR2F* (52%) was reduced to <2% in all cell models except hPCLiS 0h (9%) while for *HSPA8*, it was reduced from 65% to <3% for all except iPSC HLC (25%). Similar trends can be seen for *GOLGA8B* and *ARHGAP21* (Fig. 10*)*.

An investigation at this level revealed major changes in transcript usage for all in vitro cell models (Fig. 5). After removing the DTUs from the non-DEGs, termed as non-DEGs^{DTU-} (Fig. 3), their count and pathway coverage decreased. As for non-DEGs, a similar trend can be seen for non-DEGs^{DTU-} in terms of pathway coverage. Pathway mapping data of non-DEGs^{DTU-} for each cell model for all KEGG pathways

are also provided to investigate queries per cell model and/or pathway (Suppl. Table 3b).

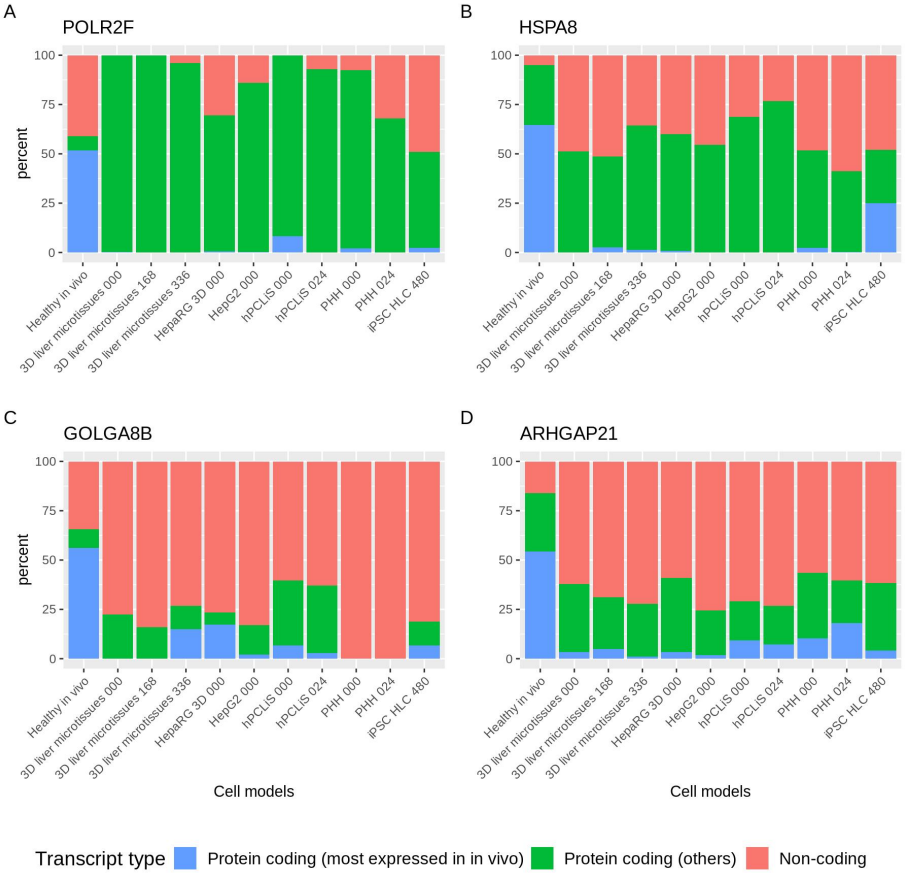


***Figure 10**: Examples of four non-DEGs that show major differential transcript usage (DTU). Transcript usage illustrated for four genes that were not differentially expressed at the gene level (non-DEGs) but had differential transcript usage (DTU). The most expressed protein coding transcript in vivo is replaced by other protein coding and/or non-coding transcripts (A) POLR2F, (B) HSPA8, (C) GOLGA8B, and (D) ARHGAP21.*

Discussion

Different *in vitro* liver cell models have been developed for studying the effects of toxic compounds in humans. In the past, these models have been evaluated time and again for specific processes and components, giving a limited overview [13, 17, 18], however, a systematic comparison of RNAseq data is not yet available. We

compared these models at baseline gene expression using RNA-Seq. While *in vivo* and *in vitro* samples were sequenced on different platforms (Illumina NovaSeq 6000® and HiSeq 2000® respectively), it is important to consider that both samples were produced using the same library preparation method, and both Illumina sequencers produced comparable results. The *in vivo* samples were sequenced with longer reads (150bp) compared to cell samples (100bp) but this cannot be expected to cause larger differences in the data as the read length higher than 50bp does not drastically impact the outcome [43, 44]. Once a read's position can be mapped unambiguously, longer reads do not add much value in a quantification-based analysis [45].

As expected, in comparison to the human *in vivo* liver, the highest pairwise spearman's correlation was shown by the non-cancerous human liver derived cell models, such as 3D liver microtissues and PHH. The cancer-derived cell models and iPSC-HLCs were still classified liver based on CellNet analysis but obtained the lowest classification scores using the human liver as reference. CellNet provides an easy and direct way to compare the cell models but it uses single-end (SE) reads for building the consensus expression profiles and GRNs[36] to accommodate more data available in public domains. However, SE sequences have poor coverage and low resolution of the 3' end of the transcripts as compared to paired-end sequences. Thus, further approaches besides CellNet are required.

Therefore, we analyzed non-DEGs to focus on similarities between the *in vivo* and *in vitro* samples. Based on the number of non-DEGs and pathway coverage, 3D liver microtissues initially showed a high similarity with the liver *in vivo* but during the cultivation period of 168 and 336h, the number of non-DEGs decreased. The largest deviation from the results obtained by CellNet was obtained with the hPCLiS and iPSC-HLC samples. Based on the results of CellNet, hPCLiS showed a high level of similarity with *in vivo* liver but using non-DEGs a relatively low resemblance was observed. On the other hand, for iPSC-HLC, CellNet predicted poor similarity but non-DEGs demonstrated a higher degree of similarity. The difference between results obtained by CellNet and analysis of non-DEGs could be due to differences in the level of lowly expressed genes which may gain more weight in the analysis of non-DEGs than in CellNet due to downsampling in CellNet [36]. Furthermore, with DEGs, the highest similarity was observed for 3D liver microtissues and PHH, and lowest for HepG2 and hPCLiS 000 for the enrichment analysis and pathway coverage.

In addition to non-DEGs and DEGs, we also explored DTU thus highlighting the genes which are not differentially expressed on the gene level but exhibited significant differential usage of isoforms on the transcript level. The change in the amount of expression of the different transcript types in the cell models provided another metric to distinguish liver similar and dissimilar cell models. The analysis of the DTU first resulted in reduced numbers of non-differentially expressed genes (non-DEGs^{DTU}) and hence pathway coverage. While globally the results of pathway coverage were similar to non-DEGs, studying the DTUs helped in identifying the genes which were differentially spliced between the *in vivo* liver and the *in vitro* system, notably by having a dominant protein coding or non-coding transcript(s). An important point worth mentioning here is that in our methodology, we used a stricter p-value cut-off in the case of DTU (0.01) because isoform mapping is known to induce a higher false-positive rate [46, 47]. The evaluation of the DTUs aid in identifying the regulation control of expression of different protein coding and non-coding transcripts and conservation of the function of the proteins which otherwise remains oblivious at the gene level, as illustrated for the four exemplary genes in Fig. 7: These four genes are responsible for the process of transcription to protein trafficking and localization. First, *POLR2F* is a component of RNA polymerases I, II, and III which plays an important role in transcription [48] while *HSPA8* which is involved in a wide variety of cellular processes and also takes care of protein folding, transport, and proteolysis [49]. *GOLGA8B* and *ARHGAP21* are responsible for maintaining the Golgi apparatus [50, 51] and were shown to be differentially expressed at the transcript level. The differential expression of these genes implies that the functions of the Golgi (modifying, sorting, and packaging of proteins for secretion) may be perturbed. The present results show that the 3D liver microtissues (0h) demonstrate a particularly high Spearman correlation, CellNet classification score, GRN status, number of non-DEGs, non-DEGs^{DTU-}, and pathway coverage. During the cultivation period, these values decrease. It should, however, be considered that cultivation periods of 168 and 336h are relatively long and it is difficult to maintain in vivo like properties for such long periods. For short term incubation of 24h PHH represent an adequate system, since the CellNet classification score, GRN status, number of non-DEGs, non-DEGs^{DTU-}, and pathway coverage remained almost unchanged during the cultivation period. Therefore, in agreement with previous studies [52, 53] cultivated primary hepatocytes seem to represent an adequate system for short term experiments to identify genome-wide expression changes. Over the incubation period, the non-DEG pathway coverage

for primary bile acid biosynthesis increased for PHH. This is in agreement with previous studies showing that isolated hepatocytes establish bile canaliculi and express bile acid excretion carriers at their apical membranes during the first 24h in culture [54, 55]. While the hPCLiS cell models exhibited lower similarity with *in vivo* liver compared to microtissues and PHH, this may be explained by the location of extraction of the tissue from the liver cancer patients.

HepG2 cells lost numerous functions compared to primary hepatocytes. Nevertheless, they are used for *in vitro* studies as they represent a relatively inexpensive, easy to handle cell line. These present a higher intermodal variability, probably because the cancer cells under uncontrolled cell division accumulate mutations over time. The same holds for HepaRG cells that still show slightly more non-DEGs than HepG2 cells but are less similar to *in vivo* liver tissue as PHH or microtissues. Several recent studies reported that HepaRG 3D models mimic *in vivo* liver [56-58] but these studies did not perform an RNA-Seq based comparison to human liver tissues.

The use of human iPSCs as a renewable source for the generation of human hepatocytes holds great promise as non-transformed hepatocytes from individuals with multiple genetic backgrounds could be generated. However, consistent with other publications [16, 31, 59], we here found that iPSC-derived hepatocyte-like cells still show major differences compared to liver tissue and primary human hepatocytes. CellNet analysis of the RNA-seq expression profiles of human-iPSC-HLCs demonstrated that the iPSC progeny shows a low CellNet classification score for the human liver. Moreover, they still share a resemblance with embryonic stem cells and exhibit some overlap with the expression profiles of the intestine and colon cells/tissue (Suppl. Table 2), as previously described in other studies [31]. Additionally, non-DEGs and non-DEG^{DTU-} were identified by comparing the mRNA profiles of the iPSC-HLCs with *in vivo* liver expression data. With around 7118 non-DEGs and 7087 non-DEG^{DTU-} iPSC-HLCs demonstrated an even higher similarity to human liver tissue than hPCLiS, HepaRG, and HepG2 but ranged clearly behind microtissues and PHH. However, when mapping onto liver pathways selected from KEGG, the iPSC HLCs showed only a relatively low pathway coverage. Taken together, the results illustrate that iPSC-derived cells performed better than the cancer models (HepG2 and HepaRG) and in some cases even better than hPCLiS as well. Though these results suggest that they exhibit some similarity to *in vivo* liver, there are still significant hurdles to overcome before iPSC-derived hepatic progeny

reach a high similarity to real hepatocytes. Different strategies to improve HLC differentiation may include chemical engineering of the culture media [60], the use of 3D organoid cultures and microfluidic systems to recreate the *in vivo* hepatocyte niche and to allow the manipulation of oxygen gradients and the delivery/removal of specific factors [61, 62]. In addition, as several TFs are highly differentially expressed between iPSC-HLCs and *in vivo* liver, another way to improve iPSC-HLCs maturation could be by up/downregulation of these misregulated TFs (Suppl. Fig. 5f) [16, 31]. It is important to consider that these results were obtained from baseline comparisons and, while analyzing or deriving hypothesis from these results, it should be kept in mind that their response to stress and/or exposure to chemicals still has to be elucidated.

Data availability

The raw data can be assessed from ENA: *In vivo* liver: PRJEB35350/ERP118386, PHH: PRJEB23590/ERP105351, iPSC-HLC: PRJEB23620/ERP105382, 3D liver microtissues: PRJEB24482/ERP106310, HepG2: PRJEB24466/ERP106294 and PRJEB24464/ERP106292, HepaRG 3D: PRJEB24487/ERP106315, and hPCLiS: PRJEB24484/ERP106312.

References

1. Cribb AE, Peyrou M, Muruganandan S, Schneider L: **The endoplasmic reticulum in xenobiotic toxicity**. *Drug Metab Rev* 2005, **37**(3):405-442.
2. Moeller TA, Shukla SJ, Xia M: **Assessment of compound hepatotoxicity using human plateable cryopreserved hepatocytes in a 1536-well-plate format**. *Assay Drug Dev Technol* 2012, **10**(1):78-87.
3. Dey N, De P, Smith BR, Leyland-Jones B: **Of mice and men: the evolution of animal welfare guidelines for cancer research**. *British Journal of Cancer* 2010, **102**(11):1553-1554.
4. Ericsson AC, Crim MJ, Franklin CL: **A Brief History of Animal Modeling**. *Missouri medicine* 2013, **110**(3):201-205.
5. Hau J: **Animal models for human diseases**. In: *Sourcebook of models for biomedical research.* Springer; 2008: 3-8.
6. Simmons D: **The use of animal models in studying genetic disease: transgenesis and induced mutation**. *Nature education* 2008, **1**(1):70.
7. DelRaso NJ: **In vitro methodologies for enhanced toxicity testing**. *Toxicol Lett* 1993, **68**(1-2):91-99.
8. Godoy P, Hewitt NJ, Albrecht U, Andersen ME, Ansari N, Bhattacharya S, Bode JG, Bolleyn J, Borner C, Boettger J: **Recent advances in 2D and 3D in vitro systems using primary hepatocytes, alternative**

hepatocyte sources and non-parenchymal liver cells and their use in investigating mechanisms of hepatotoxicity, cell signaling and ADME. *Archives of toxicology* 2013, **87**(8):1315-1530.

9. LeCluyse EL, Bullock PL, Parkinson A: **Strategies for restoration and maintenance of normal hepatic structure and function in long-term cultures of rat hepatocytes**. *Advanced Drug Delivery Reviews* 1996, **22**(1):133-186.
10. Sachinidis A, Albrecht W, Nell P, Cherianidou A, Hewitt NJ, Edlund K, Hengstler JG: **Road map for development of stem cell-based alternative test methods**. *Trends in molecular medicine* 2019.
11. Albrecht W, Kappenberg F, Brecklinghaus T, Stoeber R, Marchan R, Zhang M, Ebbert K, Kirschner H, Grinberg M, Leist M: **Prediction of human drug-induced liver injury (DILI) in relation to oral doses and blood concentrations**. *Archives of toxicology* 2019, **93**(6):1609-1637.
12. Gebhardt R, Hengstler JG, Müller D, Glöckner R, Buenning P, Laube B, Schmelzer E, Ullrich M, Utesch D, Hewitt N: **New hepatocyte in vitro systems for drug metabolism: metabolic capacity and recommendations for application in basic research and drug development, standard operation procedures**. *Drug Metab Rev* 2003, **35**(2-3):145-213.
13. Guillouzo A, Corlu A, Aninat C, Glaise D, Morel F, Guguen-Guillouzo C: **The human hepatoma HepaRG cells: a highly differentiated model for studies of liver metabolism and toxicity of xenobiotics**. *Chemico-biological interactions* 2007, **168**(1):66-73.
14. Kanebratt KP, Andersson TB: **HepaRG Cells as an in Vitro Model for Evaluation of Cytochrome P450 Induction in Humans**. *Drug metabolism and disposition* 2008, **36**(1):137-145.
15. Gu X, Albrecht W, Edlund K, Kappenberg F, Rahnenführer J, Leist M, Moritz W, Godoy P, Cadenas C, Marchan R: **Relevance of the incubation period in cytotoxicity testing with primary human hepatocytes**. *Archives of toxicology* 2018, **92**(12):3505-3515.
16. Godoy P, Schmidt-Heck W, Hellwig B, Nell P, Feuerborn D, Rahnenführer J, Kattler K, Walter J, Blüthgen N, Hengstler JG: **Assessment of stem cell differentiation based on genome-wide expression profiles**. *Philosophical Transactions of the Royal Society B: Biological Sciences* 2018, **373**(1750):20170221.
17. Wilkening S, Stahl F, Bader A: **Comparison of primary human hepatocytes and hepatoma cell line Hepg2 with regard to their biotransformation properties**. *Drug metabolism and disposition* 2003, **31**(8):1035-1042.
18. Soldatow VY, Lecluyse EL, Griffith LG, Rusyn I: **In vitro models for liver toxicity testing**. *Toxicol Res* 2013, **2**(1):23-39.
19. Van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C: **Ten years of next-generation sequencing technology**. *Trends in genetics* 2014, **30**(9):418-426.
20. Chu Y, Corey DR: **RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation**. *Nucleic Acid Therapeutics* 2012, **22**(4):271-274.
21. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nature reviews genetics* 2009, **10**(1):57.
22. Kvam VM, Liu P, Si Y: **A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data**. *American journal of botany* 2012, **99**(2):248-256.
23. Chen G, Wang C, Shi T: **Overview of available methods for diverse RNA-Seq data analyses**. *Science China Life Sciences* 2011, **54**(12):1121-1128.
24. Bae H, Monti S, Montano M, Steinberg MH, Perls TT, Sebastiani P: **Learning Bayesian Networks from Correlated Data**. *Sci Rep-Uk* 2016, **6**.
25. Cahan P, Li H, Morris SA, da Rocha EL, Daley GQ, Collins JJ: **CellNet: Network Biology Applied to Stem Cell Engineering**. *Cell* 2014, **158**(4):903-915.

26. Li E, Davidson EH: **Building Developmental Gene Regulatory Networks**. *Birth defects research Part C, Embryo today : reviews* 2009, **87**(2):123-130.
27. Carninci P: **Is sequencing enlightenment ending the dark age of the transcriptome?** *Nature methods* 2009, **6**(10):711-713.
28. Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, Fernandez Banet J, Billis K, García Girón C, Hourlier T *et al*: **The Ensembl gene annotation system**. *Database* 2016, **2016**:baw093-baw093.
29. Matoulkova E, Michalova E, Vojtesek B, Hrstka R: **The role of the 3' untranslated region in post-transcriptional regulation of protein expression in mammalian cells**. *RNA Biology* 2012, **9**(5):563-576.
30. Mayr C: **Evolution and Biological Roles of Alternative 3'UTRs**. *Trends in cell biology* 2016, **26**(3):227-237.
31. Godoy P, Schmidt-Heck W, Natarajan K, Lucendo-Villarin B, Szkolnicka D, Asplund A, Bjorquist P, Widera A, Stober R, Campos G *et al*: **Gene networks and transcription factor motifs defining the differentiation of stem cells into hepatocyte-like cells**. *J Hepatol* 2015, **63**(4):934-942.
32. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nat Methods* 2012, **9**(4):357-359.
33. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data**. *Bioinformatics* 2014, **30**(15):2114-2120.
34. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome**. *BMC Bioinformatics* 2011, **12**:323.
35. Bahin M, Noël BF, Murigneux V, Bernard C, Bastianelli L, Le Hir H, Lebreton A, Genovesio A: **ALFA: annotation landscape for aligned reads**. *BMC genomics* 2019, **20**(1):250.
36. Radley AH, Schwab RM, Tan Y, Kim J, Lo EKW, Cahan P: **Assessment of engineered cells using CellNet and RNA-seq**. *Nature Protocols* 2017, **12**:1089.
37. Cahan P, Li H, Morris SA, Lummertz da Rocha E, Daley GQ, Collins JJ: **CellNet: network biology applied to stem cell engineering**. *Cell* 2014, **158**(4):903-915.
38. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**. *Genome Biol* 2014, **15**(12):550.
39. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: **The KEGG resource for deciphering the genome**. *Nucleic Acids Res* 2004, **32**(suppl_1):D277-D280.
40. Luo W, Brouwer C: **Pathview: an R/Bioconductor package for pathway-based data integration and visualization**. *Bioinformatics* 2013, **29**(14):1830-1831.
41. Dufour J-F, Clavien P-A, Graf R, Trautwein C: **Signaling pathways in liver diseases**: Springer; 2010.
42. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z: **GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists**. *BMC Bioinformatics* 2009, **10**(1):48.
43. Chhangawala S, Rudy G, Mason CE, Rosenfeld JA: **The impact of read length on quantification of differentially expressed genes and splice junction detection**. *Genome Biol* 2015, **16**(1):131-131.
44. Rizzetto S, Eltahla AA, Lin P, Bull R, Lloyd AR, Ho JWK, Venturi V, Luciani F: **Impact of sequencing depth and read length on single cell RNA sequencing data of T cells**. *Sci Rep-Uk* 2017, **7**(1):12781.
45. Stark R, Grzelak M, Hadfield J: **RNA sequencing: the teenage years**. *Nature reviews genetics* 2019, **20**(11):631-656.
46. Rehrauer H, Opitz L, Tan G, Sieverling L, Schlapbach R: **Blind spots of quantitative RNA-seq: the limits for assessing abundance, differential expression, and isoform switching**. *BMC Bioinformatics* 2013, **14**(1):370.
47. Soneson C, Matthes KL, Nowicka M, Law CW, Robinson MD: **Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage**. *Genome Biol* 2016, **17**(1):12.

48. Kershnar E, Wu SY, Chiang CM: **Immunoaffinity purification and functional characterization of human transcription factor IIH and RNA polymerase II from clonal cell lines that conditionally express epitope-tagged subunits of the multiprotein complexes**. *J Biol Chem* 1998, **273**(51):34444-34453.
49. Stricher F, Macri C, Ruff M, Muller S: **HSPA8/HSC70 chaperone protein: structure, function, and chemical targeting**. *Autophagy* 2013, **9**(12):1937-1954.
50. Sousa S, Cabanes D, Archambaud C, Colland F, Lemichez E, Popoff M, Boisson-Dupuis S, Gouin E, Lecuit M, Legrain P *et al*: **ARHGAP10 is necessary for alpha-catenin recruitment at adherens junctions and for Listeria invasion**: Nat Cell Biol. 2005 Oct;7(10):954-60. doi: 10.1038/ncb1308. Epub 2005 Sep 25.
51. Dubois T, Paleotti O, Mironov AA, Fraisier V, Stradal TE, De Matteis MA, Franco M, Chavrier P: **Golgi-localized GAP for Cdc42 functions downstream of ARF1 to control Arp2/3 complex and F-actin dynamics**. *Nat Cell Biol* 2005, **7**(4):353-364.
52. Grinberg M, Stöber RM, Edlund K, Rempel E, Godoy P, Reif R, Widera A, Madjar K, Schmidt-Heck W, Marchan R: **Toxicogenomics directory of chemically exposed human hepatocytes**. *Archives of toxicology* 2014, **88**(12):2261-2287.
53. Grinberg M, Stöber RM, Albrecht W, Edlund K, Schug M, Godoy P, Cadenas C, Marchan R, Lampen A, Braeuning A: **Toxicogenomics directory of rat hepatotoxicants in vivo and in cultivated hepatocytes**. *Archives of toxicology* 2018, **92**(12):3517-3533.
54. Reif R, Karlsson J, Günther G, Beattie L, Wrangborg D, Hammad S, Begher-Tibbe B, Vartak A, Melega S, Kaye PM: **Bile canalicular dynamics in hepatocyte sandwich cultures**. *Archives of toxicology* 2015, **89**(10):1861-1870.
55. Godoy P, Widera A, Schmidt-Heck W, Campos G, Meyer C, Cadenas C, Reif R, Stöber R, Hammad S, Pütter L: **Gene network activity in cultivated primary hepatocytes is highly similar to diseased mammalian liver tissue**. *Archives of toxicology* 2016, **90**(10):2513-2529.
56. Takahashi Y, Hori Y, Yamamoto T, Urashima T, Ohara Y, Tanaka H: **Three-dimensional (3D) spheroid cultures improve the metabolic gene expression profiles of HepaRG cells**. *Bioscience reports* 2015:BSR20150034.
57. Ott LM, Ramachandran K, Stehno-Bittel L: **An automated multiplexed hepatotoxicity and CYP induction assay using HepaRG cells in 2D and 3D**. *SLAS DISCOVERY: Advancing Life Sciences R&D* 2017, **22**(5):614-625.
58. Ramaiahgari SC, Waidyanatha S, Dixon D, DeVito MJ, Paules RS, Ferguson SS: **From the Cover: Three-Dimensional (3D) HepaRG Spheroid Model With Physiologically Relevant Xenobiotic Metabolism Competence and Hepatocyte Functionality for Liver Toxicity Screening**. *Toxicological Sciences* 2017, **159**(1):124-136.
59. Heslop JA, Duncan SA: **The use of human pluripotent stem cells for modelling liver development and disease**. *Hepatology*, **0**(ja).
60. Siller R, Greenhough S, Naumovska E, Sullivan GJ: **Small-molecule-driven hepatocyte differentiation of human pluripotent stem cells**. *Stem Cell Reports* 2015, **4**(5):939-952.
61. Takebe T, Sekine K, Enomura M, Koike H, Kimura M, Ogaeri T, Zhang R-R, Ueno Y, Zheng Y-W, Koike N *et al*: **Vascularized and functional human liver from an iPSC-derived organ bud transplant**. *Nature* 2013, **499**:481.
62. Ong LJY, Chong LH, Jin L, Singh PK, Lee PS, Yu H, Ananthanarayanan A, Leo HL, Toh YC: **A pump-free microfluidic 3D perfusion platform for the efficient differentiation of human hepatocyte-like cells**. *Biotechnology and bioengineering* 2017, **114**(10):2360-2370.

Chapter 3

FuSe: a tool to move RNA-Seq analyses from chromosomal/gene loci to functional grouping of mRNA transcripts

Rajinder Gupta[1], Yannick Schrooders[1], Marcha Verheijen[1], Adrian Roth[2], Jos Kleinjans[1], Florian Caiment[1*]

1. Department of Toxicogenomics, School of Oncology and Developmental Biology (GROW), Maastricht University, Maastricht, The Netherlands
2. Roche Pharmaceutical Research and Early Development, Roche Innovation Center Basel, Basel, Switzerland

Published: *Bioinformatics, 2020*

DOI: https://doi.org/10.1093/bioinformatics/btaa735

Abstract

Summary

Typical RNA-Seq analyses are performed either at the gene level by summing all reads from the same locus, assuming that all transcripts from a gene make a protein or at the transcript level, assuming that each transcript displays unique function. However, these assumptions are flawed, as a gene can code for different types of transcripts and different transcripts are capable of synthesizing similar, different, or no protein. As a consequence, functional changes are not well illustrated by either gene or transcript analyses. We propose to improve RNA-Seq analyses by grouping the transcripts based on their similar functions. We developed FuSe to predict functional similarities using the primary and secondary structure of proteins. To estimate the likelihood of proteins with similar functions, FuSe computes two confidence scores: knowledge (KS) and discovery (DS) for protein pairs. Overlapping protein pairs exhibiting high confidence are grouped to form 'similar function protein groups' and expression is calculated for each functional group. The impact of using FuSe is demonstrated on *in vitro* cells exposed to paracetamol, which highlight genes responsible for cell adhesion and glycogen regulation which were earlier shown to be not differentially expressed with traditional analysis methods.

Availability: The source code is available at https://github.com/rajinder4489/FuSe. Data for APAP exposure are available in the BioStudies database (http://www.ebi.ac.uk/biostudies) under accession numbers S-HECA143, S-HECA(158), and S-HECA139

Supplementary information: Data are available at Oxford Bioinformatics online

Introduction

With the evolution of RNA sequencing (RNA-Seq), an immense amount of high-quality transcriptomics data has been generated; identifying and quantifying each gene transcript/isoform with high precision. Transcriptomics data are often studied to identify the changes in gene or transcript expression between different conditions and treatments. The ones exhibiting the highest perturbation at the lowest statistical error are then mapped to pathways and ontologies to illuminate the functional consequences of the alteration. However, the typical data analysis pipelines to assess gene expression from RNA-Seq dataset are not perfect. The expression level of a given gene is usually obtained by the summation of expression (read count) of all the different spliced variants (isoforms/transcripts) mapping the gene locus. These spliced variants are identified using the sequence identity to the genome and chromosomal locus. These isoforms can be protein coding (same or different proteins), non-coding, nonsense mediated decay, or else. Considering the level of expression of a gene as the summation of all reads from these different types of isoforms is misrepresentative as it considers all of them as coding for the same protein.

Alternatively, analyzing RNA-Seq data at the level of isoform can also be performed by keeping an individual read count for every single transcript. Keeping each isoform separated would assume that there are no functional overlaps between different transcripts. Currently, most of the tools available to quantify RNA-Seq data like RSEM [1], StringTie [2], Sailfish [3], Salmon [4], Kallisto [5], and HT-Seq [6] along with others, focus chiefly on gene or transcript (isoform) expression. Cuffdiff [7], another read counts quantifying tool, groups different transcripts from the same transcription start site (TSS) to identify genes that are differentially regulated at the transcriptional or post-transcriptional level. None of these tools captures the functional similarity of the proteins coded by different transcripts. However, we know that closely related proteins are capable of exhibiting same functions and these proteins might be derived from different genes (paralogs) or from the same gene locus via alternative RNA splicing. Different histones [8] – HS1.1, HS1.2, HS1.3, HS1.4, HS1.5 originating from same family of genes HIST1H1A-E, respectively, share functional similarities. Another well studied case is the ubiquitin-conjugating enzymes [9] – E2D1, E2D2, E2D3, E2D4 which originate from UBE2D1-4 genes, respectively.

There is no denying that functional overlap between proteins, derived from different genes and transcripts, exists and analyses focused on individual genes or transcripts would fail to translate to actual functional changes. A paradigm shift has to take place to move from gene/transcript-based to function-based analyses. To assess the importance of a given function, the actual amount of all proteins able to perform this function would need to be quantified. However, quantifying the proteins using the state of the art proteomics technologies do not allow to have the exhaustive panel of expressed proteins [10] and hence, mRNA quantification data (using RNA-Seq) is a better alternative for establishing functional analyses. Indeed, all proteins' primary structure can be predicted from their corresponding mRNA, a multitude of tools such as Translation Tool [11], EMBOSS Transeq [12] or TranslatorX [13] are developed to accomplish this task using the knowledge of codon to an amino acid relationship, translation start or ORFs (Open reading frames). Moreover, considering the limitation of the proteomics, the quantified mRNA expression from the RNA-Seq experiments can provide a surrogate evaluation of protein expression at steady state [14] and can be quantified using RNA to protein conversion factors otherwise [15].

Comparing the protein function and ontology profiles would provide the list of highly similar proteins; however, this would require a comprehensive protein-function-ontology knowledgebase which is not available. Around ~20k SwissProt and ~168k TrEMBL entries on Uniprot (date accessed: 20/04/2019) are available for humans [16]. In the absence of such information, protein structure (tertiary and quaternary) seems a reliable option, as the function is chiefly defined by the structure. The lack ofhigh-resolution structures, ~3.5k proteins with less than 1.5 Å and ~13.5k with 1.5 - 2.0 Å for humans on PDB (date accessed: 20/04/2019) [17], and the unavailability of pure state protein structures due to protein stability poses a hindrance in defining the structure-function relationships. Different artificial intelligence and machine learning approaches have been employed to predict the protein structures [18] but with limited precision and success because of multiple attraction and repulsion forces in action.

The only extensive high-quality information on the proteins available is the nucleotide sequence of their corresponding mRNAs. A comparison of these mRNAs is unsuited to find the similarity between them because of the presence and differences in intronic regions, UTRs (3' or 5'), and CDS. Moreover, comparing the nucleotide sequences does not take into account the degeneracy (redundancy) of

the genetic codon. Hence the amino acid (primary protein) sequence of these proteins is taken for comparison. The primary sequence is readily available but in the case of unknown proteins, it can be achieved from the mRNA sequences. Their comparison can illustrate the local and global sequence identities but it is not enough to predict the functional similarity of the proteins. To supplement the comparison, data on the secondary structures are added to the comparison. A specific order of these structures gives rise to supersecondary structures and these can be used in envisaging the structural and functional features of the protein [19].

Rather than defining gene expression that groups together the transcripts from the same chromosomal locus, we developed FuSe (Functional grouping of transcripts for RNA-Seq analyses) with an aim to group protein coding transcripts based on the predicted similarity of their protein function. For this, we used the available primary structure of proteins and predicted the secondary, super secondary structures, and protein families from it. For establishing the similarity from the protein primary sequence, BLAST+ [20] was used to identify sequence identity, coverage, and gaps in the alignment. While sequence identity establishes the similarity between the proteins, the coverage provides information if the two sequences match globally or locally. The gaps further help in checking the presence of any insertions or deletions and hence provide information about the alignment continuity. Interpro is an ensemble of 14 different tools developed using state of the art algorithms and knowledgebase to find and predict the domains, motifs, and protein families [21].

On the foundations of this information, two types of confidence scores: knowledge (KS) and discovery (DS), are calculated for all protein pairs. KS is stringent and predicts highly similar protein pairs whereas DS is lenient and predicts the proteins with local similarity as well. Based on the confidence score, 'similar function protein groups' (SFPGs) are formed from the overlapping protein pairs and are used for recalculating the RNA-Seq expression. To assess the approach and illustrate the changes in functional inferences between the chromosomal locus and function-based grouping, mRNA data from hepatic cell models exposed to acetaminophen (APAP or paracetamol) were used.

Methods

All the protein coding transcripts for human were downloaded from Ensembl (Homo_sapiens.GRCh38.pep.all.fa) [22]. For the workflow of FuSe refer Suppl. Fig. 1.

Data preparation

To find the similarity between the protein sequences, BLAST+ (v.2.8.0) was used and data was generated in tabular format 6 of BLAST+. Then, to find and predict the presence of structural and functional domains in the proteins, Interpro (v.5.31-70.0) was used. The data were obtained in the ".tsv" format.

The output from Interpro was a list of functional and structural domains obtained from various tools embedded in Interpro. Using in-house developed scripts, for each protein, these domains are first ordered based on their position on the amino acid sequence per tool. These ordered domains were then compared for similarities between the protein pairs. The similarity between the ordered domains was labeled for each tool per protein pair as STONM (same type, order, and number of motifs), STNM (same type and number of motifs), STM (same type of motifs), and NM (no match) (Suppl. Fig. 2). STONM defines the highest level of similarity. Another term, NP (not present), was assigned to cases where there was no prediction by an Interpro tool for at least one of the proteins in the given protein pair. Each protein pair will have one term (STONM, STNM, STM, NM, or NP) for each of the 14 tools and, from this information, a protein-protein domain profile per tool is obtained.

Protein pair confidence scores

From the BLAST+ and protein-protein domain profile comparison, the protein pair confidence scores were calculated using a scoring scheme (Suppl. methods**)**. Two types of confidence scores: DS and KS, were calculated; succinct and expanded equations are given below.

Succinct equations:

$$DS = \frac{(AIS+ACS+ AGS)*100}{Max\ AIS} \quad \text{...Eq. 1}$$

$$KS = AIS + ACS + AGS + \frac{ITCS*100}{Max\ ITCS} \quad \text{...Eq. 2}$$

where

AIS: Alignment Identity score

ACS: Alignment Coverage score

AGS: Alignment Gap score

ITCS: Sum of Interpro Tools' comparison score

Expanded equations:

$$Discovery\ score\ (DS) = \left(\begin{array}{c} \frac{\%\ Identity*Identity\ score}{100} + \\ \sum_{i=1}^{2} \frac{\left|\left(100-\frac{seq\ length}{alignment\ length}*100_i\right)\right|*\ Coverage\ score}{100} + \\ \frac{Number\ of\ gaps*Gap\ score}{100} \end{array} \right) * \frac{100}{Identity\ score} \quad \text{...Eq. 3}$$

$$Knowledge\ score\ (KS) = \left(\begin{array}{c} \frac{\%\ Identity*Identity\ score}{100} + \\ \sum_{i=1}^{2} \frac{\left|\left(100-\frac{seq\ length}{alignment\ length}*100_i\right)\right|*\ Coverage\ score}{100} + \\ \frac{Number\ of\ gaps*Gap\ score}{100} + \\ \sum_{j=1}^{14} \begin{cases} if\ comparison = STONM|STNM|STM|NM, & Interpro\ tool\ score_j \\ else\ if\ comparison = NP, & skip \end{cases} \end{array} \right) * \frac{100}{\sum_{j=1}^{14} \begin{cases} if\ comparison=STONM|STNM|STM|NM, & Max\ Interpro\ tool\ score_j \\ else\ if\ comparison=NP, & skip \end{cases}} \quad \text{...Eq. 4}$$

The DS relies only on the sequence similarity attributes obtained from sequence alignment such as identity, coverage, and gaps. While identity and coverage score have a positive value, they illustrate the similarity between the protein pair, the Gap score has a negative value and demonstrates their dissimilarity. The score obtained is then normalized to 100 using the maximum possible alignment identity score. In the case of KS, the final score is a result of sequence similarity attributes and ordered secondary structure similarity given as STONM, STNM, STM, NM, or

NP. To avoid penalizing protein with missing prediction information for one (or more) of the 14 Interpro tools (Suppl. Fig. 3), we then normalize the ITCS to the maximum possible ITCS score.

Similar function protein groups (SFPGs)

Confidence scores were calculated from all possible protein pairs as described in the previous step. To identify the proteins which are similar in function, a confidence score cutoff (CSC) was used with a default value of KS≥95. The CSC can be any positive integer ≤100. A lower CSC would result in the formation of SFPG with false positives. It is important to establish here that a given transcript can be a member of one or more SFPGs, as it can have certain a sufficient amount of similarity (above the assigned DS or KS threshold) with transcripts belonging to different SFPGs.

Calculating SFPG expression

This step is divided in two parts: first the normalization of the raw reads followed by the calculation of the SFPG expression. For the normalization, the raw read counts of the transcripts and their effective length were used as calculated by RSEM.

(1) Normalization: For the calculation of the SFPG expression, the read counts need to be both in-sample and across-samples normalized. While FPKM is in-sample normalized, it is not comparable across different samples. Additionally, while the normalized read counts generated using one of state of the art method (such as DESeq2 or, edgeR) focuses on normalizing for library depth and genes densities to compare the same transcript among different treatment groups, it does not allow the absolute comparison of transcripts of a different length. To address these concerns, we combined these two normalization approaches and generated expression which is in-sample and across-samples normalized. We first normalized the raw read counts for the transcripts using the DESeq2 default normalization method and then, using the effective length of the transcripts as given by the RSEM, converted these normalized read count into FPKM (c.f. normalized_fpkm module on FuSe's GitHub repository).

(2) Expression of SFPGs: Using the normalized FPKM and SFPGs formed in step 3, SFPG expression is then calculated. The calculation of the SFPG expression can be achieved using one of two proposed approaches available in FuSe: (a) equal distribution (ED) or (b) group size distribution (GD) available under "recal_expression" module. In the case of equal distribution, the expression of the transcript is equally divided between all the SFPGs of which it is a member (equation 5), thus giving equal importance and weight to all individual members of all SFPG. For the group size distribution, the expression of the SFPGs is based on the number of members present in each SFPG (equation 6) (Suppl. Fig. 4). If the equal distribution is used, each function, as defined by a SFPG, is given equal importance whereas group size distribution is based on the concept of genetic redundancy [23], giving higher importance to bigger groups. Group size distribution is illustrated in Suppl. Fig. 5.

$$Equal\ distribution\ =\ \sum_{i=1}^{No.of\ members\ in\ SFPG} \frac{Normalized\ FPKM\ of\ member_i}{No\ of\ SFPGs\ for\ member_i}$$

…Eq. 5

$$Group\ size\ distribution\ =$$

$$\sum_{i=1}^{No.of\ members\ in\ SFPG} \frac{No\ of\ members\ in\ current\ group}{No\ of\ members\ in\ all\ groups\ of\ member_i} * member_i\ normalized\ FPKM$$

…Eq. 6

Using the GRCh38 for humans from Ensembl, we have created the data object (bi_do; BLAST Interpro data object) which can be used for further analyses. For using the future updates to Ensembl protein sequences, BLAST+ or/and Interpro, create a new bi_do using the steps mentioned on FuSe's Github repository. The data object provided is generated using the protein coding transcripts only however, if the user intends, other types of transcripts can also be used for instance nonsense mediated decay, polymorphic pseudogene, non-stop decay, etc. refer Readme for FuSe's repository on GitHub.

Assessment of FuSe

To illustrate the significance of using FuSe, we used a RNA-Seq dataset obtained from a 3D human hepatic cell model (Primary Human Hepatocytes + Kuepfer cells Spheroids from InSphero®) exposed to APAP. Ribo-depleted libraries were generated from these cell models and sequenced on an Illumina Hiseq2000 (Suppl. Methods) at an average of 41.3 million reads per sample in 100bp paired-end (Suppl. Fig. 6). There were four sets of samples: control untreated (ConUNTR), control exposed to DMSO (ConDMSO), exposed to therapeutic dose (Ther), and exposed to toxic dose (Tox). ConUNTR and Ther had five time points: 0, 2, 8, 24, and 72 hours, and ConDMSO and Tox had four time points: 2, 8, 24, and 72 hours. Each time point had three replicates, totaling 54 samples. The therapeutic dose was calculated based on PBPK (physiologically based pharmacokinetic) modeling using human kinetic data, and the toxic dose was obtained from IC20 [24]. Data are available in the BioStudies database (http://www.ebi.ac.uk/biostudies) under accession numbers S-HECA143, S-HECA(158), S-HECA139. Reads above Q30 for ConUNTR, ConDMSO, Ther, and Tox samples constituted 85.19%, 88.43%, 94.5%, and 94.2% of all reads, respectively. FPKM for transcripts was calculated using RSEM and from its expression for SFPGs was calculated using FuSe. PCA (principal component analysis) and hierarchical clustering were done for top 500 expressed transcripts using R packages: prcomp() and hclust(), respectively, for isoform FPKM and recalculated SFPG expression to compare them. Then differentially expressed transcripts (DETs) were evaluated using Anova package in R for all dose vs control samples: ConUNTR v/s The, ConUNTR v/s Tox, ConDMSO v/s The, ConDMSO v/s Tox and Ther v/s Tox samples. Significant cutoff was set to p-value<0.01 and |log2FC|>1. Changes in the DETs between the original FPKM and recalculated expression were also established.

Results

In order to move from a loci-based transcriptomics analysis to a function-based analysis, we first need to identify all protein coding transcripts from the human genome. For this, a total of ~107k proteins were retrieved from Ensembl, of which 719 proteins were discarded which originated from transcripts annotated as nonsense mediated decay, polymorphic pseudogene, T-cell receptor genes, and immunoglobulin genes. From the remaining transcripts, protein pairs were formed

and two confidence scores (KS and DS) were calculated to estimate the likelihood of similar functions. Taken individually, both scores are used to make the protein pairs and a confidence score cutoff (CSC, here used: 85, 90, and 95 for both KS and DS) is introduced to discard the protein pairs with low similarity. While both DS and KS show a steep increase in the number of pairs with the lowering of CSC, a steep increase for DS can be seen (Fig. 1A). The protein pairs can be divided into four categories depending on the origin of the transcripts, namely: same gene, same gene family, different gene, or undefined (Fig. 1B). A considerable amount of protein pairs originated from the same or different gene families at KS and DS ≥95. Moreover, a surge can be witnessed with decreasing CSC.

With the protein pairs at different CSC, SFPGs are formed and a similar trend of increase in the group size (Fig. 1C), the sum of group sizes per CSC and maximum group size for a SFPG (Suppl. Fig. 5(A)) with decreasing CSC can be seen. The median for the group size per CSC remains low (Suppl. Fig. 5(B)), implying that most groups are small. The increase in group size for the biggest group per CSC was also apparent (Suppl. Fig. 5(C)). For KS and DS at ≥95, the largest group size was 84 and 111 which rose to 216 and 229 at ≥85, respectively.

To evaluate the impact of SFPG on biological interpretation, we applied FuSe on an *in vitro* transcriptomics dataset obtained from a hepatic cell model exposed to different doses (therapeutic and toxic) of APAP for variable time duration and corresponding controls (untreated and DMSO). The dataset was analyzed by locus-based and our function-based method. For this, SFPGs at KS≥95 were taken to particularly consider the highly similar protein pairs predicted using maximum available knowledge. The expression for the SFPGs was calculated from the normalized FPKM using both methods, namely, '*ED*' and '*GD*' (c.f. section 2.4, Calculating SFPG expression). Differences in the recalculated expression from *ED* and *GD* are discussed in the Supplementary Results, Suppl. Fig. 7. The primary analyses shown in this paper makes use of the recalculated expression obtained using the '*GD*' method to give a higher importance to bigger SFPGs, implying that important biological processes are conserved. The term recalculated expression from here on designate the recalculated expression from '*GD*'. The normalized FPKM and recalculated expression (GD) from FuSe were then compared.

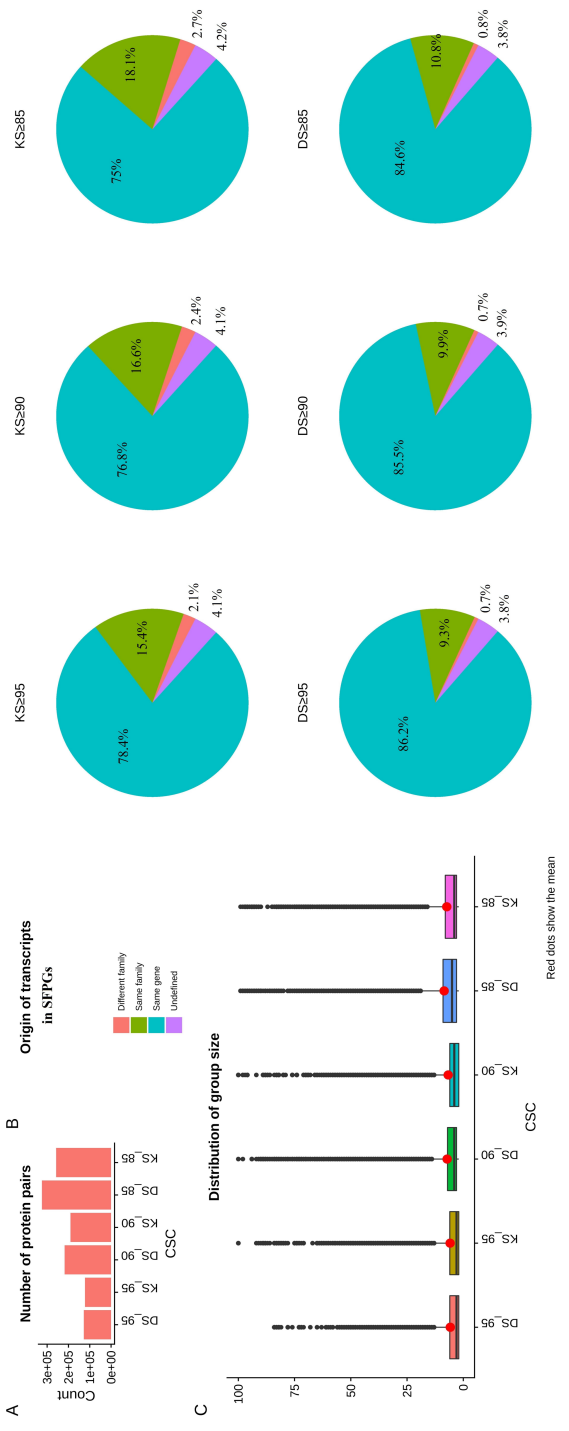


***Figure 1**: Characterization of protein pairs and similar function protein groups (SFPGs) formed for KS and DS ≥ 85, 90, and 95. (A) The number of protein pairs (B) The origin of the transcripts that make the protein pairs. These can be derived from the same gene, same gene family, different gene family, or unknown; it is determined based on the gene names. The genes which are assigned numeric identifiers, sometimes with version numbers, are categorized as undefined. (C) The distribution of the number of groups. The mean (given as red dots) can be seen increasing as the CSC is decreased but the median stays comparatively lower.*

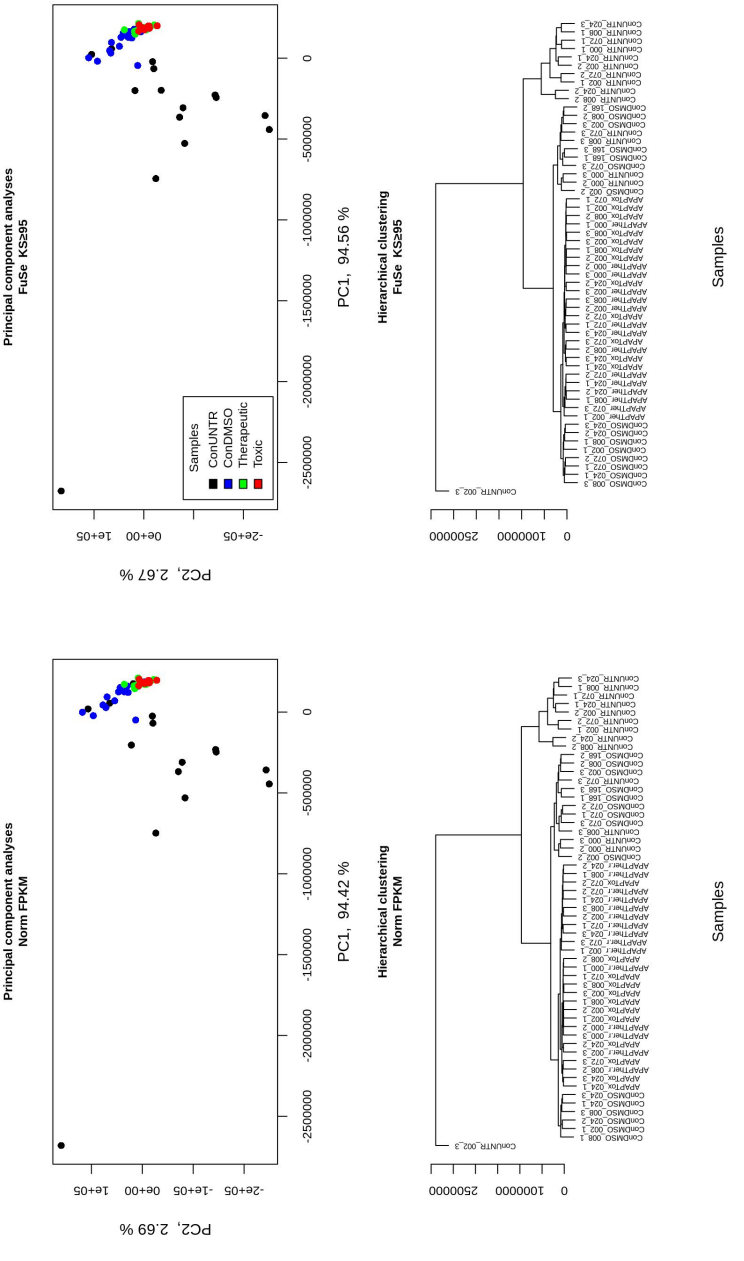


Figure 2: *The PCA bi-plots and clustering of the original FPKM and recalculated expression at KS≥95. Little to negligible effect was observed on PCA plots. From the clustering, it can be seen that the biggest source of variation is dose and is conserved after the recalculation. The samples are named as 'treatment_timepoint_replicate'.*

PCA bi-plots and clustering (Fig. 2, Suppl. Fig. 8 (A-E)) were done to show that FuSe preserves the inter-sample variation and global profile of the samples. The highest source of variation between the samples was dose and it was conserved.

Furthermore, differentially expressed transcripts (DETs) were calculated for the FPKM and recalculated expression obtained using FuSe, individually for all control v/s dose samples. Lesser number of DETs were observed after applying FuSe (Fig. 3, Suppl. Table 1). We also observed many transcripts that were significantly differentially expressed in FPKM but non-significant after using FuSe and vice-versa (Fig. 3, blue bars and green bars, respectively). With loci based analyses, the differences were computed at the transcript level while using FuSe the changes at the functional level were captured. Using FuSe, the expression levels were correctly quantified as a result of the expression of other similar function proteins. In the case of ConUNTR v/s Tox and ConDMSO v/s Tox, this is illustrated by protein coding transcripts from many genes responsible in cell adhesion and tight junction such as *PKP2-201* (Plakophilin-2), *CHCHD3-203* (MICOS complex subunit), *CHCHD3-201* (MICOS complex subunit MIC19), *IMMT-205* (MICOS complex subunit MIC60), *AGRN-201* (Agrin), *WDR1-205* (WD repeat-containing protein 1), *CTNNA1-243* (Catenin alpha-1), *ZBTB33-202* (Transcriptional regulator Kaiso), *ASPH-201* and *ASPH-207* (Aspartyl/asparaginyl beta-hydroxylase). All these transcripts, not considered differentially expressed by the standard analysis method, were found significantly affected by a toxic dose of APAP after applying FuSe.

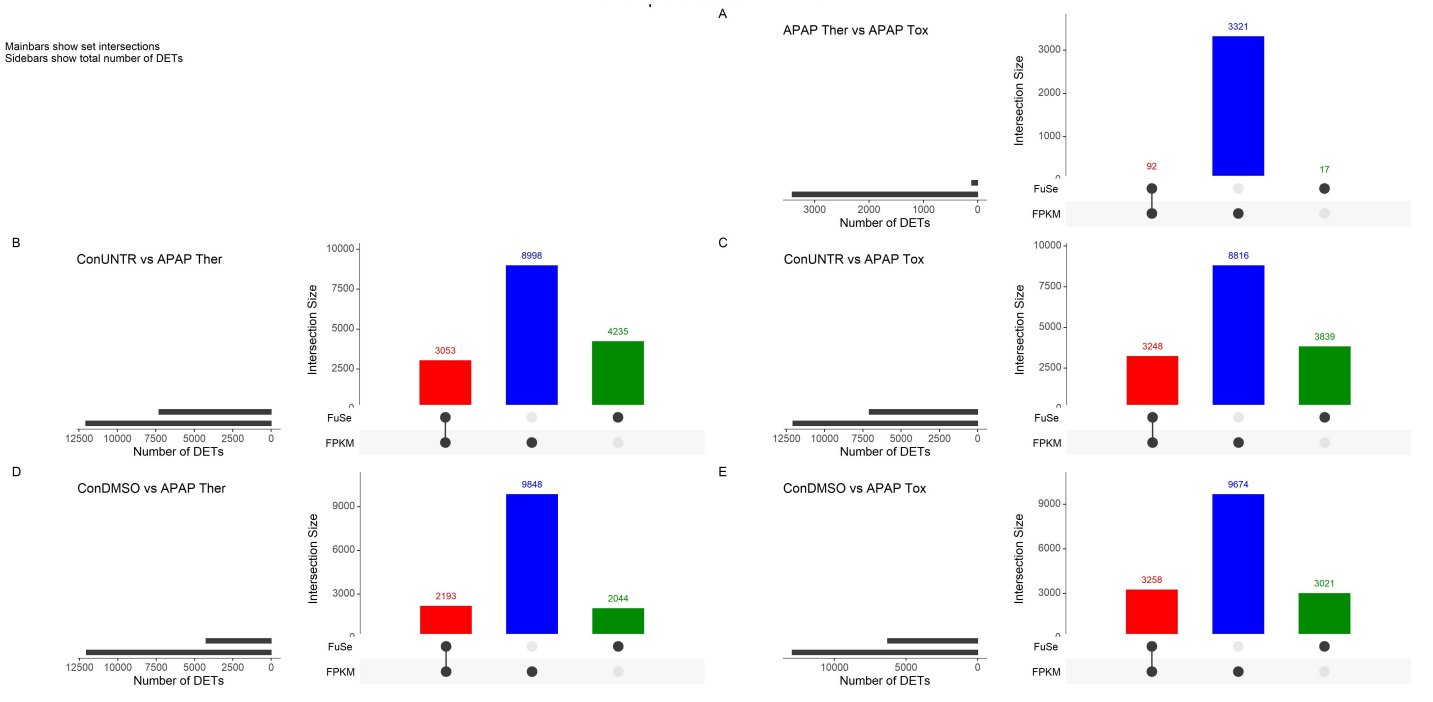


Figure 3: *Comparison of DETs. Change in DETs (basemean>10 and pval<0.01 and |log2FC|>1) obtained from FPKM and FuSe expression. The DETs were calculated using DESeq2 and the plots were made using UpSetR package [25] in R. (A)Ther v/s Tox, (B) ConUNTR v/s Ther, (C) ConUNTR v/s Tox, (D) ConDMSO v/s Ther, and (E) ConDMSO v/s Tox. The red bars show overlapping DETs, blue and green bars show exclusive DETs from original FPKM and recalculated expression, respectively.*

More importantly, some transcripts displaying a significant regulation (up or down) using the conventional analysis method were found to be significantly regulated in the opposite direction after using FuSe (Fig. 4, Suppl. Fig. 9). The highest number of switches was seen for ConUNTR v/s Ther. A total of 79 unique transcripts changed their direction of regulation (from upregulated to downregulated and vice-versa), and 3727 changed from differentially expressed to not differentially expressed (and vice-versa) across all comparisons. As an example, *PPP1R14B-203* (Protein phosphatase 1 regulatory subunit 14B) and *GBE1-205* (1,4-alpha-glucan-branching enzyme), which are protein coding transcripts, demonstrated a change in the direction of regulation. While for *PPP1R14B-203*, the change could be seen for ConDMSO v/s Ther and ConDMSO v/s Tox, for GBE1-205, the switch was witnessed for ConUNTR v/s Tox and ConDMSO v/s Tox.

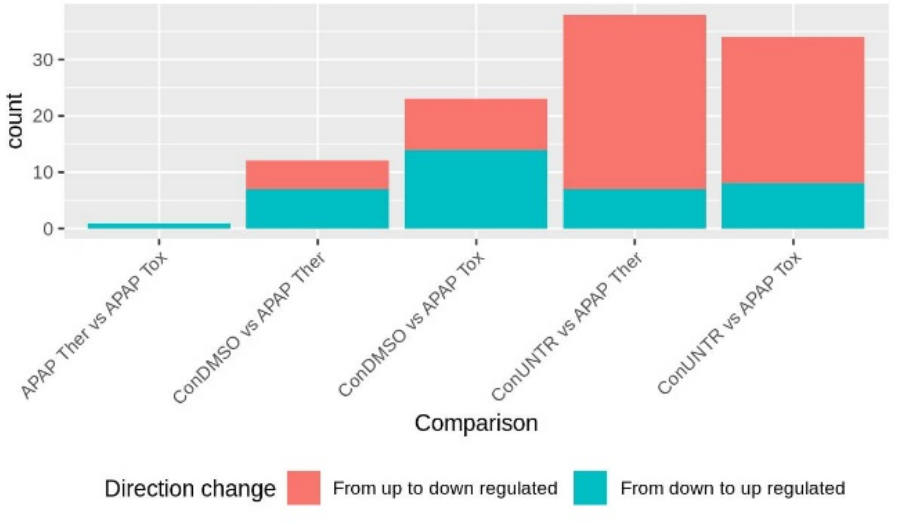


***Figure 4**: Number of DETs changing the direction of regulation. The figure illustrates the number of transcripts for which the change in the direction of regulation of the DETs (basemean>10 and pval<0.01 and |log2FC|>1) was observed. The comparison was made for DETs from standard FPKM and recalculated FPKM analyzed using FuSe.*

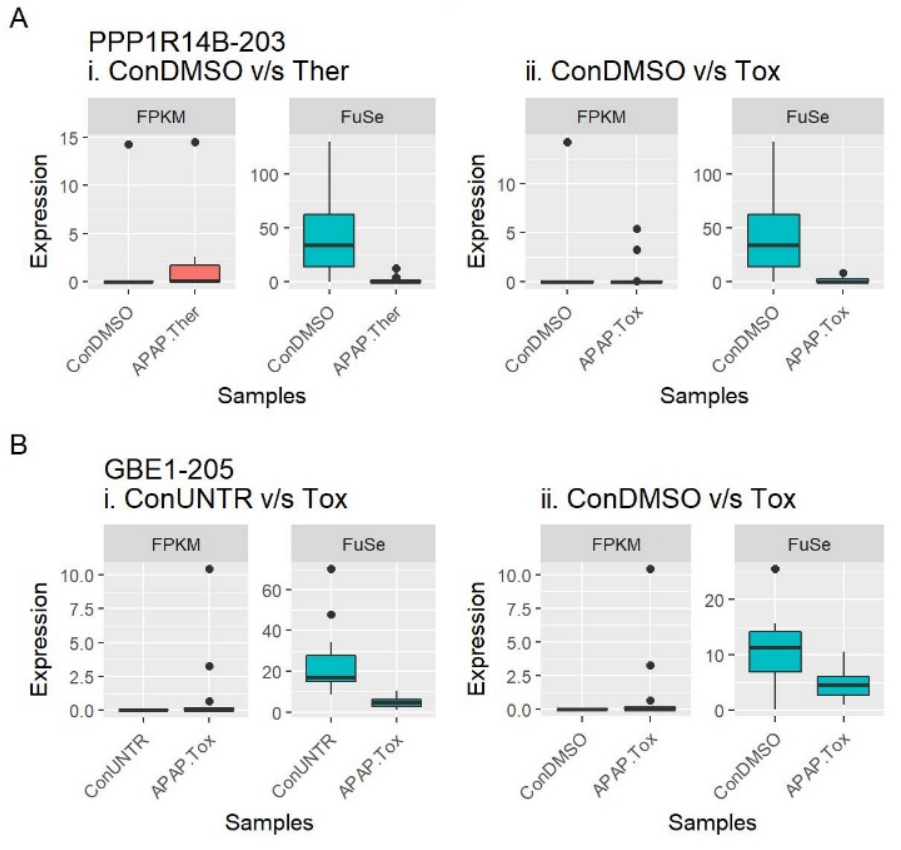


Figure 5*: Effect of using FuSe. Several DETs changed their direction of regulation (up or down regulated); here illustrated using two cases. (A) PPP1R14B-203 was upregulated with the standard method but, after the functional grouping based analyses, it was shown to be down regulated for (i) ConDMSO v/s APAP Ther and (ii) ConDMSO v/s APAP Tox. (B) Similarly in the case of GBE1-205, for (i) ConUNTR v/s APAP Tox and (ii) ConDMSO v/s APAP Tox, it was exhibited as downregulated regulated after applying FuSe.*

FuSe also demonstrated how gene expression based analyses sometimes lead to incorrect results. As there are multiple different protein coding transcripts for a gene, to compare the expression of the gene to the transcripts we selected the longest protein coding transcripts for the comparison. Moreover, for the APAP study, there were ~59% genes where the longest protein coding transcript was the highest expressed transcript; making them a suitable candidate for the comparison. The differences between the gene and SFPG expression can be illustrated here using two cases, *POLR2J2* (RNA polymerase II subunit J2) from ConUNTR v/s Tox and *UBE2D4* (Ubiquitin-conjugating enzyme E2 D4) from ConUNTR v/s Ther (Fig. 6).

While the gene expression of *POLR2J2* is contributed by two transcripts (both protein coding), the expression of SFPG (*POLR2J2-202*) is attributed by four similar protein coding transcripts (KS = 98.02). The expression of the longest protein coding isoform follows the pattern of the gene expression, however, the SFPG expression shows the opposite. For *UBE2D4*, the gene expression comprises of expression of 11 transcripts (one retained intron, one processed transcript, seven nonsense mediated decay, and two protein coding), the SFPG expression of the longest protein coding transcript from the gene (*UBE2D4-201*) is contributed by 18 similar protein coding transcripts (KS = 96.05). The longest protein coding transcript follows the pattern of SFPG though the magnitude is much higher as illustrated by SFPG.

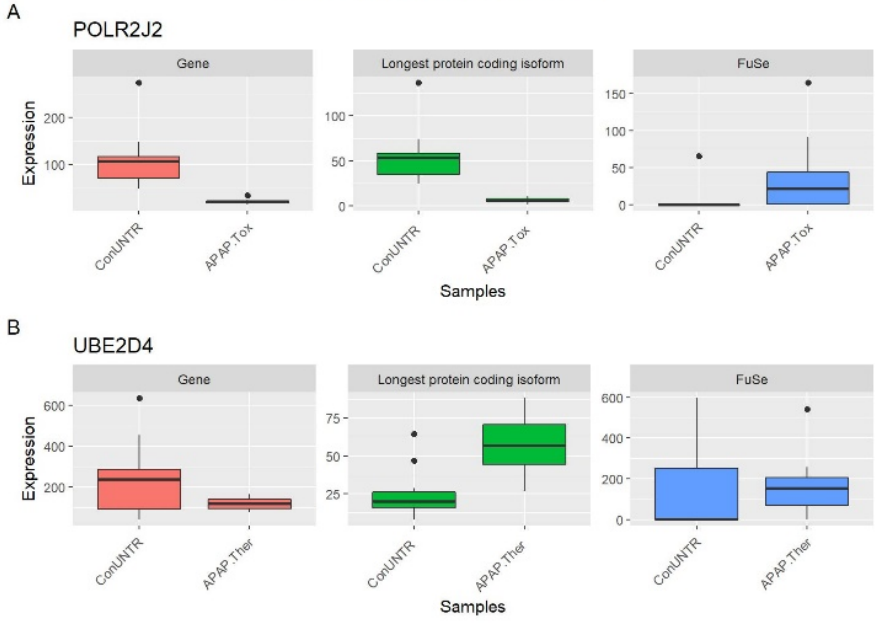


***Figure 6**: Expression of the gene, the longest protein coding isoform, and SFPG from FuSe. The typical gene expression analysis is obtained by summing all the transcripts (protein coding and non-coding) from the locus whereas in our proposed functional grouping, only the transcripts making the same protein are grouped and expression is calculated for the SFPG. The longest protein coding isoform from the gene was chosen as a representative of the gene to be compared directly with SFPG. (A) POLR2J2; the gene expression is the result of two protein coding transcripts whereas the SFPG is constituted by four similar protein coding transcripts. (B) UBE2D4; the gene expression is the result of 11 transcripts (two protein coding and nine non-coding transcripts) whereas the SFPG is constituted by 18 similar protein coding transcripts. Similar function proteins can arise from the same or different genes.*

Discussion

Biological research aims to find the functional properties and changes in the biological system, which makes us question the very fabric of our current RNA-Seq analyses strategies that focuses on changes in individual genes or transcripts. While the information of each gene or transcript is informative, however at the system level, contribution by each element should be viewed in terms of functional change. Elevating the analyses of the RNA-Seq to the functional level will increase our understanding of the biological systems and their underlying processes. Due to the limitations in quantifying the whole panel of expressed proteins and limited knowledge of protein tertiary and quaternary structure, and functions identifying the proteins having similar functions is challenging. Here, we focused on the primary and secondary structure of the proteins to establish their functional similarity.

The hypothesis that similar primary sequence and secondary structures make the proteins possessing similar functional properties can be challenged at various levels, however, it is an acceptable hypothesis [26] and holds true for most cases as shown by studies of homologs and paralogs [27]. Protein trees calculated from sequence similarity often have the same topologies as those calculated from structural similarity. While some cases might be overlooked for instance a point mutation, which can result in a different conformation of the protein, they will not have a significant influence on the sequence alignment and secondary structure prediction (if not present in the prediction region). Moreover, other external factors such as molecular crowding or macro-molecular environment cannot be accounted for while taking into account only the primary and secondary structures.

Even though primary and secondary structures exhibit some limitations to predict similar proteins, they allow us to group similar function proteins and form SFPGs. FuSe uses amino acid sequences of the proteins thus eliminating all non-coding transcripts from the analyses. While many non-coding transcripts have been associated with a specific function, not enough data are available at this stage to allow a generic *in silico* grouping of non-coding RNA sharing a similar biological function. Moreover, the transcripts annotated as nonsense mediated decay, polymorphic pseudogene, T-cell receptor genes, and immunoglobulin genes were removed from the formation of SFPGs. Nonsense mediated decay transcripts were removed because they are destined to be decayed before entering the ribosomal

machinery for protein synthesis. In the case of the polymorphic pseudogenes, they have lost their functional properties over evolution and may result in non-functional proteins. T-cell receptor and immunoglobulin genes are very selective for their targets and even if they share high similarity among them, they have different affinities for their targets and hence cannot be considered as a similar functional entity. Even though these transcript types are removed from the creation of SFPGs, they are retained with their original expression in the recalculated expression.

The SFPGs are hence formed of only the protein coding transcripts. The similarity of the proteins is established using the two score types: DS and KS. DS makes an over-prediction because it relies only on the amino acid sequence, providing little information on the final structure and hence function(s) of the protein. KS is more conservative and takes into account other available knowledge to establish similarity. The lenient nature of DS makes it a powerful tool to find novel protein pairs whereas KS, being stringent, under-estimates and limits the ballooning of the groups by keeping only highly similar proteins together. DS can be used in finding the local similarities in proteins and thus allows the discovery of a subset of common functions between two proteins whereas KS thrives for global functional similarity.

It is worth noting here that a SFPG is made for each protein coding transcript that has other similar protein coding transcripts, in order to preserve the specialized protein functions. It resulted in redundancy, as some transcripts can be a member of multiple SFPGs and these SFPGs might then be semi- or fully- overlapping. If such SFPGs were merged, it would result in the formation of false-positive SFPGs (decreased specificity) and the consecutive loss of some specialized protein functions. As some transcripts were shared between many groups, for the calculation of expression of the SFPGs, two methods are made available. First, equally dividing the expression of member transcripts between all the overlapping SFPGs and the other where the expression is divided between the SFPGs based on the group size of each SFPG, giving higher expression to bigger groups. While the first method is conservative giving equal importance to all SFPGs, the second is biased towards the bigger SFPGs, establishing that the important functions are more preserved. The quantification of the SFPG expression using FuSe requires in-sample normalized data because it needs to compare the transcripts within the sample for its calculation. Moreover, the expression should also be normalized

across samples to compare it across samples. We calculated the FPKM from the normalized expression which was obtained using DESeq2. As a consequence, SFPGs are advised to be studied in relative analysis comparing a given SFPG between two biological conditions rather than for evaluating their absolute expression level among the different SFPGs.

The results from the comparison of FPKM and expression of SFPGs using FuSe from the hepatic cell model established that the changes in the expression of the transcripts acquired using FuSe do not change the overall look of the samples, though the changes at the level of SFPGs were apparent and pointed towards different functional inferences. For instance, genes responsible for cell adhesion and tight junctions that were initially shown to be not differentially expressed. However, the application of FuSe completely changes the biological interpretation of this signal, confirming the documented APAP effects [28, 29].

There were also cases of transcripts that were differentially expressed in the opposite direction after correction using FuSe, e.g. *PPP1R14B-203* (ConDMSO versus Ther) and *GBE1-205* (ConUNTR versus Tox) (Fig. 4). The correction by FuSe reversed the direction of perturbation and hence completely changed the inferences drawn from the results. *PPP1R14B* is responsible for inhibition of *PPP1CA*, which is involved in different processes such as cell division, regulation of glycogen metabolism, muscle contractility, and protein synthesis via dephosphorylation [30, 31]. For the therapeutic dose of APAP, the upregulation of *PPP1R14B* as shown by FPKM based differential expression would imply all these processes to be inhibited. Similarly, *GBE1* was shown to be upregulated under APAP toxic dose, implying that the glycogen accumulation in the liver has increased. However, APAP is known to induce glycogen depletion and is considered as one of the early biomarkers of acetaminophen-induced hepatotoxicity [32]. Using FuSe, the GBE1 function is shown to be down regulated. The use of SPFGs also demonstrated why studying gene expression to attain differentially expressed genes can be miss leading, as in the case of *UBE2D4* (gene expression: down regulated; SFPG expression: upregulated) and *POLR2J2* (gene expression: upregulated; SFPG expression: down regulated). *UBE2D4* is involved in ubiquitination [33] and *POLR2J2* is an important component of RNA polymerase II. The downregulation of *POLR2J2* would result in a decrease in transcription while the upregulation of *UBE2D4* implies more ubiquitination leading to increased protein degradation. This suggests a decrease in protein levels in the cell, and hence

the disruption of cell processes, which is consistent with the knowledge on APAP overdose.

FuSe showed how moving from loci-based to function-based analyses changed the inferences derived from the RNA-Seq data. It illustrated functional changes that could not be captured using the conventional RNA-data analyses. Moreover, FuSe is forward compatible and new data that will be available in the future for transcripts' protein sequences and secondary structures can be integrated into the analysis by following the steps mentioned for 'Creating your own BLAST Interpro data object' under the Methodology section. Lastly, the transcripts coding for the same protein and originating from overlapping chromosomal locations but annotated to different genes have to be studied further to understand the processes and signals responsible for guiding different genes to make similar proteins. In the future, we will look to fine-tune the calculation of CSC using other inherent features of the proteins such as molecular weight, charge, electrophoretic properties, active sites, or hydrophobic-hydrophilic properties. Furthermore, to integrate all these features, new pipelines and algorithms will be investigated and developed.

References

1. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome**. *BMC Bioinformatics* 2011, **12**:323.
2. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL: **StringTie enables improved reconstruction of a transcriptome from RNA-seq reads**. *Nature biotechnology* 2015, **33**(3):290.
3. Patro R, Mount SM, Kingsford C: **Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms**. *Nature biotechnology* 2014, **32**(5):462.
4. Patro R, Duggal G, Kingsford C: **Salmon: accurate, versatile and ultrafast quantification from RNA-seq data using lightweight-alignment**. *BioRxiv* 2015:021592.
5. Bray NL, Pimentel H, Melsted P, Pachter L: **Near-optimal probabilistic RNA-seq quantification**. *Nature biotechnology* 2016, **34**(5):525-527.
6. Anders S, Pyl PT, Huber W: **HTSeq—a Python framework to work with high-throughput sequencing data**. *Bioinformatics* 2014, **31**(2):166-169.
7. Trapnell C: **Cuffdiff (v7)**.
8. Mariño-Ramírez L, Kann MG, Shoemaker BA, Landsman D: **Histone structure and nucleosome stability**. *Expert Rev Proteomics* 2005, **2**(5):719-729.
9. Hegde AN: **Ubiquitin-Proteasome System and Plasticity**. In: *Encyclopedia of Neuroscience*. Edited by Squire LR. Oxford: Academic Press; 2009: 1-9.
10. Pappireddi N, Martin L, Wühr M: **A Review on Quantitative Multiplexed Proteomics**. *Chembiochem* 2019, **20**(10):1210-1224.

11. Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, De Castro E, Duvaud S, Flegel V, Fortier A, Gasteiger E: **ExPASy: SIB bioinformatics resource portal**. *Nucleic Acids Res* 2012, **40**(W1):W597-W603.
12. McWilliam H, Li W, Uludag M, Squizzato S, Park YM, Buso N, Cowley AP, Lopez R: **Analysis tool web services from the EMBL-EBI**. *Nucleic Acids Res* 2013, **41**(W1):W597-W600.
13. Abascal F, Zardoya R, Telford MJ: **TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations**. *Nucleic Acids Res* 2010, **38**(suppl_2):W7-W13.
14. Liu Y, Beyer A, Aebersold R: **On the Dependency of Cellular Protein Levels on mRNA Abundance**. *Cell* 2016, **165**(3):535-550.
15. Koussounadis A, Langdon SP, Um IH, Harrison DJ, Smith VA: **Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system**. *Sci Rep-Uk* 2015, **5**:10775-10775.
16. The_UniProt_Consortium: **UniProt: a hub for protein information**. *Nucleic Acids Res* 2015, **43**(Database issue):27.
17. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic Acids Res* 2000, **28**(1):235-242.
18. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A: **Critical assessment of methods of protein structure prediction (CASP)-Round XII**. *Proteins* 2018, **1**:7-15.
19. Pelley JW: **3 - Protein Structure and Function**. In: *Elsevier's Integrated Biochemistry.* Edited by Pelley JW. Philadelphia: Mosby; 2007: 19-28.
20. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: **BLAST+: architecture and applications**. *BMC Bioinformatics* 2009, **10**(1):421.
21. Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, Brown SD, Chang H-Y, El-Gebali S, Fraser MI: **InterPro in 2019: improving coverage, classification and access to protein sequence annotations**. *Nucleic Acids Res* 2018, **47**(D1):D351-D360.
22. Frankish A, Vullo A, Zadissa A, Yates A, Thormann A, Parker A, Gall A, Moore B, Walts B, Aken BL *et al*: **Ensembl 2018**. *Nucleic Acids Res* 2017, **46**(D1):D754-D761.
23. Nowak MA, Boerlijst MC, Cooke J, Smith JM: **Evolution of genetic redundancy**. *Nature* 1997, **388**(6638):167-171.
24. Kuepfer L, Clayton O, Thiel C, Cordes H, Nudischer R, Blank LM, Baier V, Heymans S, Caiment F, Roth A *et al*: **A model-based assay design to reproduce in vivo patterns of acute drug-induced toxicity**. *Archives of toxicology* 2018, **92**(1):553-555.
25. Conway JR, Lex A, Gehlenborg N: **UpSetR: an R package for the visualization of intersecting sets and their properties**. *Bioinformatics* 2017, **33**(18):2938-2940.
26. Gong H, Rose GD: **Does secondary structure determine tertiary structure in proteins?** *Proteins* 2005, **61**(2):338-343.
27. Jensen RA: **Orthologs and paralogs - we need to get it right**. *Genome Biol* 2001, **2**(8):INTERACTIONS1002-INTERACTIONS1002.
28. Gamal W, Treskes P, Samuel K, Sullivan GJ, Siller R, Srsen V, Morgan K, Bryans A, Kozlowska A, Koulovasilopoulos A *et al*: **Low-dose acetaminophen induces early disruption of cell-cell tight junctions in human hepatic cells and mouse liver**. *Sci Rep-Uk* 2017, **7**:37541-37541.
29. Blair JB, Hinton DE, Miller MR: **Morphological changes in trout hepatocytes exposed to acetaminophen**. *Marine Environmental Research* 1989, **28**(1):357-361.
30. Song H, Pu J, Wang L, Wu L, Xiao J, Liu Q, Chen J, Zhang M, Liu Y, Ni M *et al*: **ATG16L1 phosphorylation is oppositely regulated by CSNK2/casein kinase 2 and PPP1/protein phosphatase 1 which determines the fate of cardiomyocytes during hypoxia/reoxygenation**. *Autophagy* 2015, **11**(8):1308-1325.

31. Mi J, Guo C, Brautigan DL, Larner JM: **Protein phosphatase-1alpha regulates centrosome splitting through Nek2**. *Cancer Res* 2007, **67**(3):1082-1089.

32. Gautam R, Chandrasekar B, Deobagkar-Lele M, Rakshit S, Kumar B. N V, Umapathy S, Nandi D: **Identification of Early Biomarkers during Acetaminophen-Induced Hepatotoxicity by Fourier Transform Infrared Microspectroscopy**. *PloS one* 2012, **7**(9):e45521.

33. David Y, Ziv T, Admon A, Navon A: **The E2 ubiquitin-conjugating enzymes direct polyubiquitination to preferred lysines**. *The Journal of Biological Chemistry* 2010, **285**(12):8595-8604.

Chapter 4

Identifying novel transcript biomarkers for hepatocellular carcinoma (HCC) using RNA-Seq datasets and machine learning

Rajinder Gupta[1], Jos Kleinjans[1], Florian Caiment[1*]

1. Department of Toxicogenomics, School of Oncology and Developmental Biology (GROW), Maastricht University, Maastricht, The Netherlands

Revision: BMC Cancer

Abstract

Background

Hepatocellular carcinoma (HCC) is one of the leading causes of cancer death in the world owing to limitations in its prognosis. The current prognosis approaches include radiological examination and detection of serum biomarkers, however, both have limited efficiency and are ineffective in early prognosis. Due to such limitations, we propose to use RNA-Seq data for evaluating putative higher accuracy biomarkers at the transcript level that could help in early prognosis.

Method

To identify such potential transcript biomarkers, RNA-Seq data for healthy liver and various HCC cell models were subjected to five different machine learning algorithms: random forest, K-nearest neighbor, Naïve Bayes, support vector machine, and neural networks. Various metrics, namely sensitivity, specificity, MCC, informedness, and AUC-ROC (except for support vector machine) were evaluated. The algorithms that produced the highest values for all metrics were chosen to extract the top features that were subjected to recursive feature elimination. Through recursive feature elimination. the least number of features were obtained to differentiate between the healthy and HCC cell models.

Results

From the metrics used, it is demonstrated that the efficiency of the known protein biomarkers for HCC is comparatively lower than complete transcriptomics data. Among the different machine learning algorithms, random forest and support vector machine demonstrated the best performance. Using recursive feature elimination on top features of random forest and support vector machine three transcripts were selected that had an accuracy of 0.97 and kappa of 0.93. Of the three transcripts, two were protein coding (PARP2-202 and SPON2-203) and one was a non-coding transcript (CYREN-211). Lastly, we demonstrated that these three selected transcripts outperformed randomly taken three transcripts (15000 combinations), hence were not chance findings, and could then be an interesting candidate for new HCC biomarker development.

Conclusion

Using RNA-Seq data combined with machine learning approaches can aid in finding novel transcript biomarkers. The three biomarkers identified: PARP2-202, SPON2-203, and CYREN-211, presented the highest accuracy among all other transcripts in differentiating the healthy and HCC cell models. The machine learning pipeline developed in this study can be used for any RNA-Seq dataset to find novel transcript biomarkers.

Code: www.github.com/rajindeer4489/ML_biomarkers

Introduction

The liver, one of the largest organ in the body, performs various important functions, such as filtering harmful substances from the blood to be then excreted from the body, producing bile to help in the digestion of fats from food, or storing glycogen (sugar) that will be used for energy. Due to its continuous exposure to harmful substances, it is prone to the amplitude of diseases which can eventually cause liver failure and/or liver cancer. Cirrhosis, long term infection with hepatitis B virus, and hepatitis C virus, alcoholic liver disease, and nonalcoholic fatty liver disease (NAFLD) are leading risk factors for primary liver cancer [1]. Moreover, cancer can develop in the liver at any stage in the progression of various liver diseases. As published in independent reports by World Health Organization (WHO) [2] and the US Center for Disease Control and Prevention (CDC) [3], liver cancer is among the top causes of cancer death worldwide, of which hepatocellular carcinoma (HCC) is the most common type of primary liver cancer, accounting for ~80% liver cancers.

Reducing the global burden of HCC is, therefore, a primary concern and it can be achieved by improving early detection and management [4]. Currently, the employed prognosis for HCC includes radiological examinations and assessment of serum markers. Radiological examinations are limited for early diagnosis as the performance of the imaging techniques begins to degrade substantially below a lesion size of 2 cm and have only modest accuracy below a lesion size of 1 cm [5]. In the case of biomarkers, currently, there are ~20 biomarkers (Table 1) in research, and out of these only α-fetoprotein (alpha-fetoprotein or AFP) has a clinical application; even though it is ineffective for detecting early lesions [1, 6-8]. Of the other markers used in research, none have reached the standard level of clinical practice so far [6, 9]. However, in various studies, it has also been demonstrated that a combination of different biomarkers provides higher accuracy in predicting HCC [10-15].

***Table 1**: Currently used serum biomarkers in the prognosis of hepatocellular carcinoma (HCC).*

Used as	Biomarker(s)	Name	Comments
Individual biomarkers	AFP[10]	Alpha-fetoprotein	Increased, a sign of liver cancer
	DCP[10]	des-gamma-carboxy prothrombin	Increased, a sign of liver cancer
	GPC3[16]	Glypican-3	GPC3 is overexpressed in HCC
	GP73[17]	Golgi glycoprotein 73	High expression of GP73 in primary HCC
	MDK[18]	Midkine	Overexpressed in tumors
	OPN[19]	Osteopontin	Overexpressed
	SCCA[14]	Squamous cell carcinoma antigen	SCCA1, SCCA2 overexpressed
	ANXA2[20]	Annexin A2	Increased in HCC
	Annexin A7[21]	Annexin A7	Increased expression inhibits HCC lymph node metastasis
	CD44[22]	Cluster Differentiation 44	Increased
	CD90[22]	Cluster Differentiation 90	Increased
	CD133[23]	Cluster Differentiation 133 or prominin-1	CD133 protein expression levels of HCC in both the cytoplasm and nucleus were significantly higher than adjacent normal liver tissue.

	EpCAM[24]	Epithelial cell adhesion molecule	Tumor size, intrahepatic metastasis, and EpCAM positivity were associated with tumor recurrence
	TGF-β (1,2,3)[25]	Transforming growth factor beta	Highly activated
	FGF[26]	Fibroblast growth factor	Expression was only detected in the liver tissues of patients with chronic hepatitis type C and HCC
	HGF/SF[27]	Hepatocyte growth factor receptor	HGFA and Matriptase convert pro-HGF/SF to mature HGF/SF
Combination of biomarkers	AFP, AFP-L3, DCP [10]	Alpha-fetoprotein, Lens culinaris agglutinin-reactive fraction of alpha-fetoprotein, des-gamma-carboxy prothrombin	Increased, a sign of liver cancer
	CK19, GPC3, AFP [11]	Cytokeratin 19, Glypican-3, Alpha-fetoprotein	GPC3 with CK19 and AFP
	GPC3, HSP70, GS [12]	Glypican 3, Heat shock protein 70, Glutamine synthetase	All increased, show a better diagnosis
	TLN1, MDK [13]	Talin-1, Midkine	Talin-1 decreased, MDK increased in serum
	SCCA-AFP [14]	Squamous cell carcinoma antigen, Alpha-fetoprotein	Overexpressed
	HIF-1α, VEGF (A-D) [15]	Hypoxia-inducible factor-1α, vascular endothelial growth factor	HIF-1α and VEGF showed higher expression

Though the combinations of various biomarkers are better predictors than the individual biomarkers, sensitivity or specificity is still low for all biomarker combinations [10-15]. While proteins are the major functional element, the corresponding transcripts can be an easier surrogate to detect and quantify. The cancer-specific mRNAs can leak into the serum as a result of passive processes (such as necrosis) and active processes (such as tumor cell apoptosis and active release in microvesicles by tumor cells) [28-31]. Though non-invasive, the lack of transcriptomics data for circulating cell-free mRNAs for HCC poses a limitation in undertaking a comprehensive *in silico* study to find novel biomarkers in serum. Only one study was found where the extracellular mRNAs for three HCC cell models, namely HepG2, Huh7, and immortalized normal liver PH5CH cells were profiled [32]. On the other hand, exhaustive transcriptomics data is available for HCC tissue/cell models (c.f. Methods) and hence, we concentrated on such data to find novel HCC biomarkers.

Using RNA-Sequencing (RNA-Seq), the whole transcriptome can be quantified. Moreover, different types of transcripts (protein coding and non-coding) can also be identified. Most transcriptomics analyses focus on gene expression by aggregating the expression of all transcripts for the given gene. However, in this study, we will focus on the transcripts because alternative-splicing defects in cancer are well documented [33-35] and dysregulation of splicing variants' expression has recently emerged as a novel cancer hallmark [35]. Moreover, using the RNA-Seq data at the transcript level will also allow us to investigate the potency of non-coding transcripts to be used as biomarkers.

Machine learning (ML) is a multidisciplinary field that makes use of computer science, artificial intelligence, computational statistics, and information theory to build algorithms that learn from existing data and make predictions on new data [36]. It has found application in diverse domains of biomedicine, including, but not limited to, image analysis [37], cancer prediction from heterogeneous data [38], robust phenotyping [39], gene discovery [40], differential network analysis [41], biomarker discovery [42], and transcriptional regulated genes [43]. The application of machine learning for the biomarker discovery from the RNA-Seq data is mainly focused on genes, however, recent studies have demonstrated that transcript based analyses outperformed gene-based analyses using ML [44, 45]. To assess if transcript biomarkers have better prediction accuracy, we analyzed various HCC cell models and healthy liver RNA-Seq data. Several HCC cell models were taken for

this study (Table 2) to ascertain that their biological heterogeneity is accounted for while building the ML models. Various ML algorithms, namely random forest (RF), K-nearest neighbors (KNN), support vector machines (SVM), Naïve Bayes (NB), and Neural networks (NNET), which are extensively used in the field of biomedicine, were applied to build the models and identify novel putative transcript biomarkers for HCC.

From the transcriptomics data, three datasets were assembled: all transcripts, protein coding only and non-coding only. The goal of making these three datasets was to see if one of them provides a better prediction. Consecutively, the efficiency of the known protein biomarkers (Table 1) was also assessed by taking the transcripts for their corresponding genes. The mapped genes also comprised of protein coding and non-coding transcripts and they were also made into three datasets (as given above). The results from the complete transcriptomics data and known protein biomarkers (for all datasets) were compared to establish which dataset(s) performs better.

Methodology

The overview of the methodology is presented in Figure 1 and detailed steps are given below.

1. Data collection

a. HCC cell models: The list of all HCC human cell models was obtained from Cellosaurus [46] (Suppl. Table 1).

b. RNA-Seq data: Using the names and synonyms of these cell models, RNA-Seq datasets were searched on the European Nucleotide Archive (ENA) and were filtered for baseline expression, instrument model (Illumina HiSeq 2000 or HiSeq 2500 or NovaSeq 6000) and paired-end library layout (Table 2). The samples were also taken from the Horizon 2020 EU-ToxRisk project, as listed in Table 2.

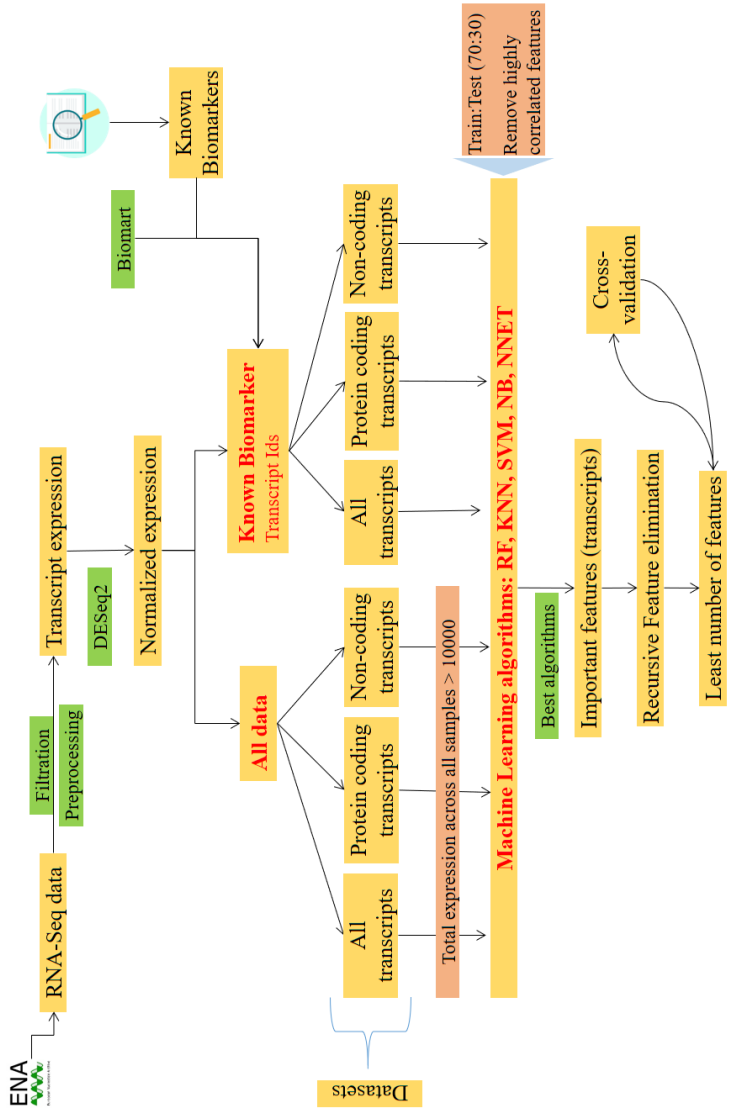


***Figure 1**: Overview of the workflow: Steps followed to find the least number of features (transcripts) required to identify the transcriptomics biomarkers. Various machine learning algorithms were used, namely random forest (RF), K-nearest neighbor (KNN), Naïve Bayes (NB), support vector machine (SVM), and neural networks (NNET).*

***Table 2**: HCC cell models and healthy liver samples were taken for this study from various studies.*

ENA		Instrument	Cell model	Type	Number of replicates
Study Id	Run accession				
PRJDB2882	DRR018792	HiSeq 2500	Huh7.5.1	HCC	1
PRJEB27210	ERR2619174, ERR2619175, ERR2619176, ERR2619177	HiSeq 2500	Hep3B	HCC	4
	ERR2619178, ERR2619179, ERR2619180, ERR2619181		HepG2	HCC	4
	ERR2619182, ERR2619183, ERR2619184, ERR2619185		HuH-7	HCC	4
PRJEB27210	ERR2619186, ERR2619187, ERR2619188, ERR2619189, ERR2619190, ERR2619191	HiSeq 2500	PHH	Healthy liver	6
PRJNA357266	SRR5104155	HiSeq 2500	LM3	HCC	1
PRJNA386625	SRR5576264, SRR5576288	HiSeq 2500	HepaRG	HCC	2
PRJNA523380	SRR8615310	HiSeq 2500	SNU-398	HCC	1
	SRR8615311		SNU-387	HCC	1
	SRR8615387		Li-7	HCC	1
	SRR8615471		SNU-878	HCC	1
	SRR8615472		SNU-886	HCC	1
	SRR8615483		JHH-1	HCC	1
	SRR8615650		SNU-475	HCC	1
	SRR8615654		SNU-423	HCC	1
	SRR8615655		SNU-449	HCC	1
	SRR8615661		HuH-7	HCC	1
	SRR8615664		HuH-1	HCC	1

	SRR8615682		SK-HEP-1	HCC	1
	SRR8615914		JHH-7	HCC	1
	SRR8615918		JHH-2	HCC	1
	SRR8615919		JHH-4	HCC	1
	SRR8615920		JHH-5	HCC	1
	SRR8615921		JHH-6	HCC	1
	SRR8615932		SNU-182	HCC	1
	SRR8615968		PLC/PRF/5	HCC	1
	SRR8616023		SNU-761	HCC	1
	SRR8616130		Hep 3B2.1-7	HCC	1
	SRR8616135		HLF	HCC	1
PRJNA206422	SRR873426	HiSeq 2000	HKCI-1	HCC	1
	SRR873427		HKCI-4	HCC	1
	SRR873428		HKCI-7	HCC	1
	SRR873429		HKCI-9	HCC	1
	SRR873430		HKCI-11	HCC	1
	SRR873836		HKCI-5B	HCC	1
EU-ToxRisk	NA	NovaSeq 6000	Healthy *in vivo* liver	Healthy liver	24*
		HiSeq 2500	Liver microtissues 3D	Healthy liver	9
			Primary human hepatocytes (PHH)	Healthy liver	11**
			Human precision-cut liver slices from HCC patients (hPCLiS)	HCC	4
			HepaRG 3D	HCC	4

			HepG2	HCC	7

*There were a total of 27 samples but three samples from children or infants were removed

**There were a total of 12 replicates for PHH, one was removed for low library depth during filtration for quality.

c. Known biomarkers: Concurrently, a list of all known biomarkers for HCC was collected through an exhaustive literature review (Table 1). These biomarkers were mapped to their corresponding Ensembl gene ids using Biomart and manual curation. In instances where there was more than one gene mapping to the protein biomarker, all instances were taken. For all the Ensembl genes that were mapped to the biomarkers, all of them had multiple isoforms/transcripts, comprising of both protein coding and non-coding transcripts.

2. Data preprocessing: The raw RNA-Seq data (fastq files) were first trimmed of their adapter sequences using Trimmomatic [47], mapped onto the human genome (version 84) from Ensembl [48] using Bowtie2 [49], and quantified using RSEM [50]. Isoform read counts were then normalized for different studies using DESeq2 [51].

3. Machine learning:

a. Preparing different datasets: We analyzed the known protein biomarkers and complete data (named as all data) separately. Furthermore, the transcriptomics data consists of protein coding and non-coding transcripts and it provided the opportunity to investigate the efficiency of different types of transcripts in identifying healthy and HCC cell models. We made three datasets, namely all transcripts (protein coding and non-coding), protein coding only, and non-coding only for both – all data and known protein biomarkers (Fig. 1).

b. Machine learning algorithms: On these six datasets (Fig. 1), machine learning algorithms from the caret package in R [52] were applied. We used five different algorithms, namely random forest (RF), K-nearest neighbors (KNN), support vector machines (SVM), Naïve Bayes (NB), and Neural networks (NNET) with ten-fold cross-validation for ten times. All further steps are applied to all six datasets individually. The seed was fixed to have reproducible results.

The data was first divided into 70:30 for training and testing, respectively. A separate validation set was not created because we used k-fold cross-validation to tune the model's hyper-parameters. In the case of datasets (all transcripts, protein

coding only, and non-coding only) from all data, all transcripts that had a total expression for all samples below 10000 were removed. This expression filter was applied to take into account only the highly expressed transcripts. However, in the case of known biomarkers, no such filter was used since we wanted to retain all information. Furthermore, using the 'findCorrelation' feature form the Caret library, highly correlated transcripts (>0.75) were identified and removed, except one (the first, a random transcript). Each algorithm's performance is assessed on all datasets by evaluating various metrics, namely sensitivity, specificity, accuracy, Matthew's correlation coefficient (MCC), and informedness (equations 1-4) using R library 'MLeval' [53] (Table 3). Additionally, the time taken by each algorithm to run is also provided.

Based on the results from these metrics, the best algorithm and dataset were selected and the top 20 important features (transcripts) were extracted using "varImp" from the Caret library. Then to find the minimum set of features to differentiate between healthy and HCC cell models, "RFE" (Recursive Feature Elimination) from the Caret library was applied using the method cross-validation (CV).

$$Sensitivity\ or\ TPR\ = \frac{TP}{TP+FN} \quad \text{...Equation 1}$$

$$Specificity\ or\ TNR\ = \frac{TN}{TN+FP} \quad \text{...Equation 2}$$

$$MCC\ = \frac{TP.TN-FP.FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad \text{...Equation 3}$$

$$Informedness = Sensitivity\ +\ Specificity - 1 \quad \text{...Equation 4}$$

where
TP is true positive
TN is true negative
FP is false positive
FN is false negative
MCC is Mathew's correlation coefficient

4. Re-training the model: The features (transcripts) selected using RFE were used to train the final model. Taking these features, exhaustive k-fold cross-validation was run by setting the repeats to 100 and number to 10; implying 1000 instances will be evaluated.

5. Chance findings: There were a total of ~200k transcripts and to establish that the features (transcripts) selected using RFE were not chance findings, 15000 iterations were performed taking three random transcripts out of the highly expressed transcripts to compare their prediction accuracy. The results from randomly taken transcripts were compared to the selected features (transcripts from RFE).

Results

To obtain an exhaustive list of all HCC *in vitro* cell models, Cellosaurus [46] was used (accessed on 27/08/2019). It houses data for 250 HCC cell models for humans (Suppl. Table 1). RNA-Seq data for all 250 cell models were searched on ENA using the application programming interface (API), taking the data generated using Illumina's HiSeq platforms or newer and library layout as paired-end. Furthermore, it was manually checked if the data were obtained at baseline. A total of 51 samples from 6 studies comprising of 33 cell models from ENA passed the filters and manual curation (Table 1). Samples from the EU-ToxRisk project were also taken; healthy *in vivo* liver (24 samples) and all other samples (32 samples from 5 cell models) were sequenced on NovaSeq 6000 and HiSeq 2500, respectively (Table 1).

The samples' quality was assessed using FastQC, and it was observed that all samples passed the "Per base sequence quality" metric. However, one sample (PHH_024_1) did not pass the library size filter and was discarded. The samples passing the filters were then processed and the transcript expression was normalized using DESeq2 for different studies.

We first investigated the expression patterns of the known biomarkers at the transcript level to see if the protein coding transcripts demonstrate a similar expression pattern as known protein biomarkers. Each gene can have multiple protein coding transcripts, only the ones mapped to manually annotated and reviewed Uniprot identifiers were considered and their expression pattern was examined (Suppl. Figure 1). VEGFA-223, HSP90AB1-203, FGF5-201, ANXA7-201, and SPP1-201 were the most down-regulated and CD44-206, HSP90AB1-201, SPP1-202, ANXA2-202, and CD44-209 were the most upregulated transcripts.

We then investigated the accuracy of the known biomarkers (all three datasets, namely all transcripts, protein coding only, non-coding only) and all data (all three datasets), in predicting the correct labels for the cell models. We focused only on

highly expressed transcripts and hence, to remove the lowly expressed ones, an expression filter was introduced (total expression across all samples > 10000 reads) (Table 3). However, in the case of known biomarkers, no such filter was used because we wanted to preserve any information, if present, held by even the lowly expressed transcripts. Furthermore, all transcripts having a high correlation (>0.75) were discarded to remove redundancy except the first (random) transcript in the list. To the remaining transcripts in each dataset, ML algorithms were applied, individually. While KNN and SVM were the fastest to run (a few seconds), NNET took the longest time for all datasets (most for all data-all transcripts: ~19 hours 44 minutes) (Table 3).

***Table 3**: Number of transcripts after steps of filtration and time to run ML algorithms on them.*

		Datasets					
		Known protein biomarkers			**All data**		
Steps		All transcripts	Protein coding	Non-coding	All transcripts	Protein coding	Non-coding
Number of transcripts after expression filter; biomarkers no filter, all data > 10000		410	262	149	16173	13688	2724
Number of highly correlated features (transcripts); correlation cutoff >0.75		177	98	37	12047	9866	1970
Number of transcripts after removing highly correlated features		234	165	113	4127	3823	755
Time to run (in seconds)	RF	10.77	8.09	6.44	196.25	169.31	32.60
	NB	12.34	9.38	6.63	297.81	280.27	46.05

	KNN	1.03	1.10	1.11	5.63	5.62	1.78
	SVM	2.25	1,07	1.05	7.51	7.48	2.72
	NNET	72.37	35.84	20.12	71044.53	56114.75	3125.74

The results obtained from the algorithms show that the area under the curve-receiver operating characteristics (AUC-ROC) values were the highest for RF and the lowest for KNN, across all datasets (Fig. 2). AUC-ROC values for SVM cannot be obtained because it is a discrete classifier. For other metrics (sensitivity, specificity, informedness, and MCC) for all datasets, SVM illustrated the highest values (Fig. 3). In the case of known biomarkers, RF demonstrated high values comparable to SVM in some cases for all datasets. NB also illustrated high values for all metrics for all data-all transcripts. We were also interested to see if protein coding or non-coding individually could give a better prediction. However, it was noted that predictions were less accurate when using them separately, as compared to all transcripts. The confidence intervals for sensitivity and specificity were the smallest in the case of all data-all transcripts for all algorithms and particularly for RF and NB (Fig. 4).

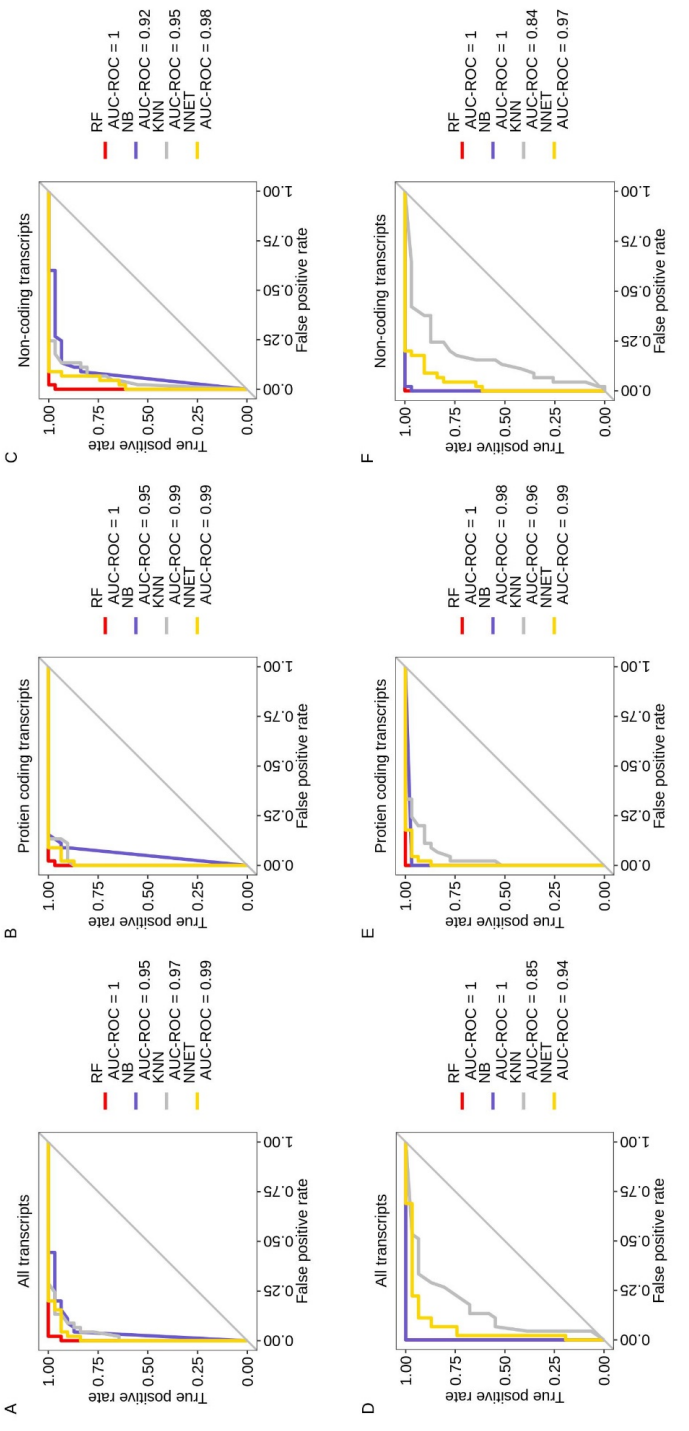


Figure 2: *AUC-ROC: The area under the ROC (receiver operating characteristic) curve for all datasets analyzed using different machine learning algorithms, namely K-nearest neighbors (KNN), Naïve Bayes (NB), neural network (NNET), and random forest (RF). Known biomarker (A: all transcripts, B: protein coding transcripts, and C: non-coding transcripts) and all data (D: all transcripts, E: protein coding transcripts, and F: non-coding transcripts).*

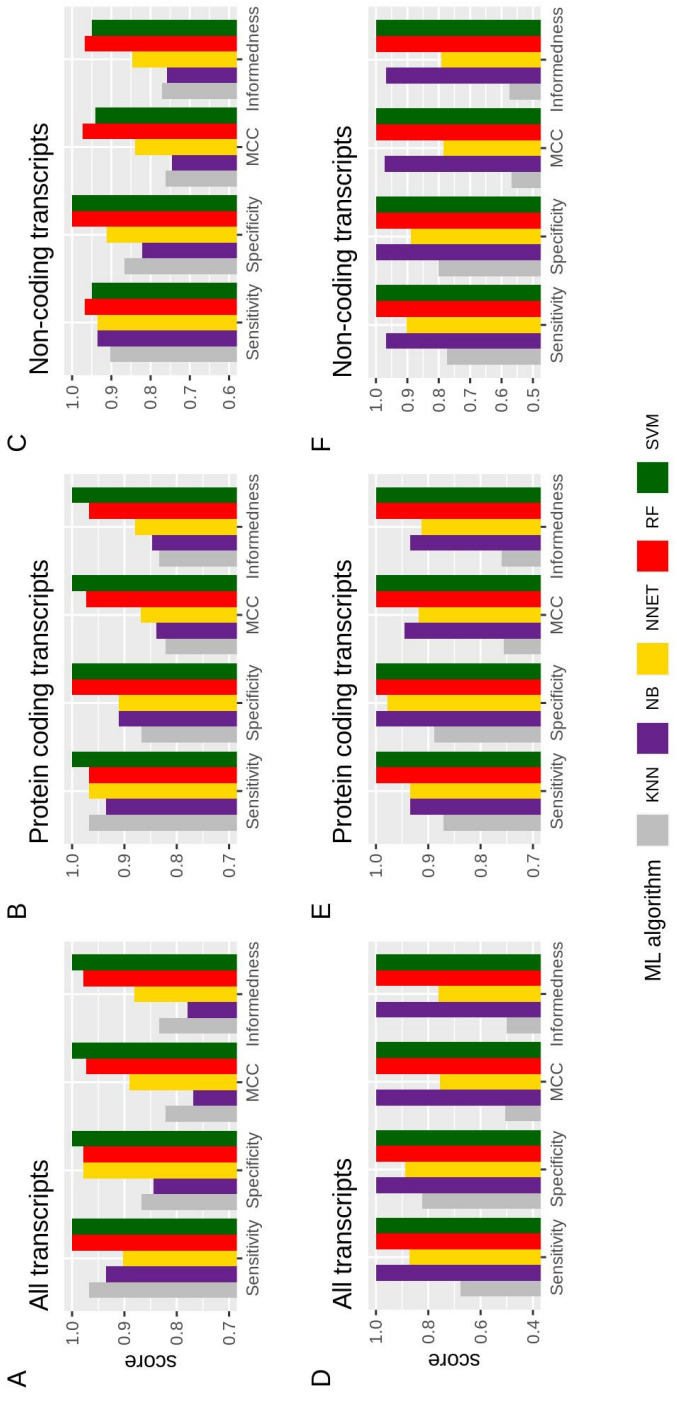


Figure 3: *Machine learning (ML) metrics values. The values for different metrics calculated for all datasets using different machine learning algorithms, namely K-nearest neighbors (KNN), Naïve Bayes (NB), neural network (NNET), random forest (RF), and support vector machine (SVM). Known biomarker (A: all transcripts, B: protein coding transcripts, and C: non-coding transcripts) and all data (D: all transcripts, E: protein coding transcripts, and F: non-coding transcripts)*

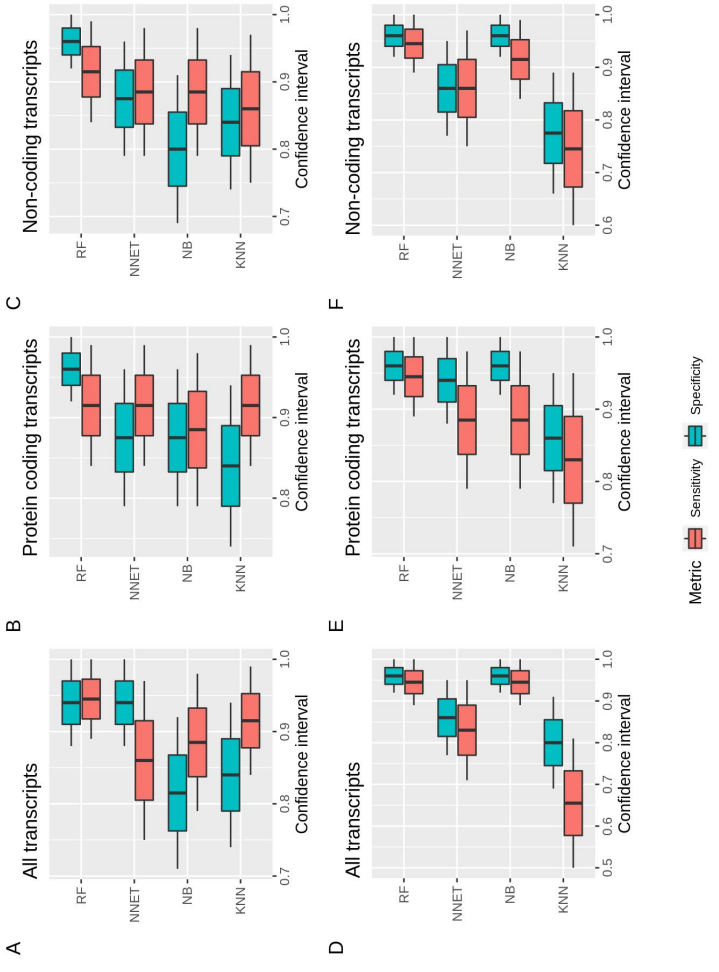


***Figure 4**: Confidence interval: Confidence interval for sensitivity and specificity across all machine learning algorithms, namely K-nearest neighbors (KNN), Naive Bayes (NB), neural network (NNET), and random forest (RF). Known biomarker (A: all transcripts, B: protein coding transcripts, and C: non-coding transcripts) and all data (D: all transcripts, E: protein coding transcripts, and F: non-coding transcripts).*

Based on the values of different metrics used to assess the performance of the algorithms on various datasets, RF and SVM performed the best for all datasets; primarily for all transcripts, protein coding transcripts, and non-coding transcripts datasets for all data. To further get the least number of features required to differentiate between the healthy and HCC cell models, the top 20 important features (transcripts) from RF and SVM when applied to all data-all transcripts were taken (Fig. 5A). There was a total of 32 unique features (transcripts), with an overlap of eight features between the two algorithms (Suppl. Figure 2). Furthermore, recursive feature elimination (RFE) was applied to this list to extract the least

number of features required to differentiate between healthy and HCC samples. With the application RFE, three features (transcripts) were identified (Fig. 5B), namely PARP2-202 (protein coding transcript), SPON2-203 (protein coding transcript), and CYREN-211 (non-coding transcript) with an accuracy of 0.97 and kappa of 0.93. These three transcripts were present in both algorithm's top important features. While PARP2-202 was upregulated (log2 fold change: 2.368), SPON2-203 and CYREN-211 were both down-regulated (-5.421 and -2.771, respectively) (Fig. 5C).

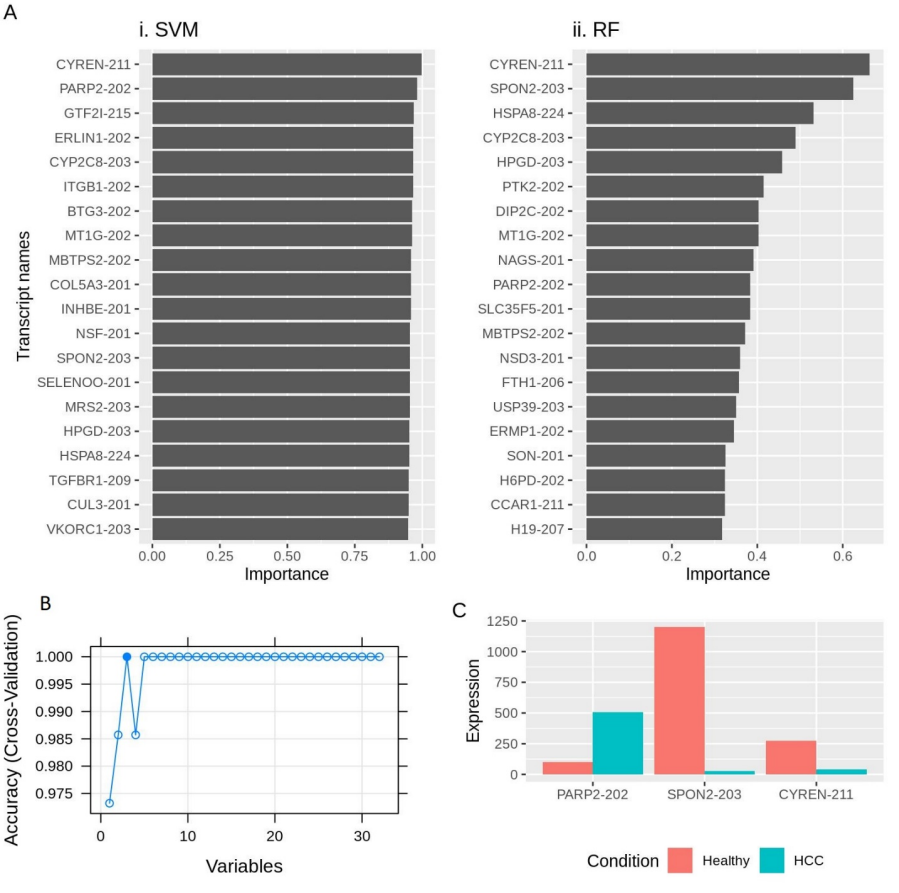


***Figure 5**: (A) Top 20 important features extracted from all data-all transcripts dataset obtained using (i) SVM and (ii) RF. (B) Recursive feature elimination (RFE) used with the top 20 features from (A) to extract a list of least number of features required to differentiate between healthy and HCC cell models. Three features were selected having an accuracy of 0.97 and kappa of 0.93 (C) Average expression of three features (transcripts), across healthy liver and HCC cell models, chosen in (B).*

A direct relationship of these genes (or transcripts) could not be established to HCC through literature review. However, an investigation of the gene ontology terms (biological process) obtained using DAVID [54] highlighted that CYREN-211 is involved in double-strand break repair via non-homologous end joining (GO:0006303) and PARP2-202 had a known function in DNA repair (GO:0006281), base-excision repair (GO:0006284) and DNA ligation involved in DNA repair (GO:0051103). In the case of SPON2-203, multiple ontologies for immune responses were obtained – GO:0002448 (mast cell mediated immunity), GO:0008228 (opsonization), GO:0032755 (positive regulation of interleukin-6 production), GO:0032760 (positive regulation of tumor necrosis factor production), GO:0043152 (induction of bacterial agglutination), GO:0045087 (innate immune response), GO:0050832 (defense response to fungus), GO:0051607 (defense response to virus), GO:0060907 (positive regulation of macrophage cytokine production), GO:0071222 (cellular response to lipopolysaccharide), GO:0001530 (lipopolysaccharide binding), and GO:0003823 (antigen binding).

To assess its strength, the model was re-trained using these three features but with an increased number of cross-validations (repeats = 100, number = 10; implying 1000 iterations). High values for all metrics were observed with RF and SVM (sensitivity: 0.968 and 0.944 (RF and SVM), specificity: 1 and 1, MCC: 0.973 and 0.936, informedness: 0.968 and 0.944, and AUC-ROC: 0.99 (RF only)). Moreover, the confidence interval for sensitivity and specificity in the case of RF was 0.84-0.99 and 0.92-1, respectively. Finally, to establish that these transcripts (PARP2-202, SPON2-203, and CYREN-211) were not chance findings, random combinations of three transcripts (highly expressed) were made and their efficiency was assessed and compared to the three transcripts selected using RFE. It was observed that out of 15000 combinations created, none of the combinations exhibited higher or equal values for the metrics for RF and only 0.12% cases (18 cases) demonstrated higher or equal value for the metrics in the case of SVM (Suppl. Table 2).

Discussion

Hepatocellular carcinoma (HCC) has a huge global burden and the challenge lies primarily in its early detection owing to the limited accuracy of serum biomarkers and inefficiency of radiological examinations. With advancements made in machine learning over the last few years, we investigated if it can assist in finding better

biomarkers for HCC. We took RNA-Seq data from HCC and healthy liver cell models and used various machine learning algorithms to highlight key features that can differentiate between the healthy and HCC cell models with high accuracy. A set of three transcripts were identified, namely PARP2-202, SPON2-203, and CYREN-211; proposed as novel putative transcript biomarkers.

Though widely studied, RNA-Seq data for HCC at baseline is not abundantly available. Out of 250 HCC cell models listed in Cellosaurus, data could only be obtained for 33 cell models. Many studies were discarded in the process of selection due to single-end library layout, low coverage, exposure to drugs various treatments, and insufficient metadata. For the 33 cell models taken in this study, 28 had only one replicate. This could have been a limiting factor if these were to be analyzed per cell model, however, in this study the focus was on HCC and all cell models were combined to define the transcriptome profile of HCC. Using the transcriptome profile, the cell type and/or condition (healthy/disease/treatment) can then be accurately assessed [55] and then comparing these profiles, distinct features for these profiles can be established.

For HCC, many biomarkers are extensively studied (Table 1), AFP being one of the most studied biomarker. Although these biomarkers have been established through studies of serum, most of them are predominantly secreted by the liver [56]. In an attempt to compare the efficiency of these known biomarkers and all data with respect to their ability to discriminate between the healthy and HCC cell models, we observed that all data out-performed known biomarkers' datasets. The comparatively lower accuracy obtained using known biomarkers can be attributed to fewer features (transcripts) in the dataset. While all data constituted of ~200k transcripts, known biomarkers amounted for ~400 features only. The transcriptomics data also provided an opportunity to investigate if protein coding or non-coding transcripts could individually be enough to classify healthy and HCC cell models. A loss of information can be witnessed in both instances compared to both types of transcripts taken together (all transcripts datasets) in case of known biomarkers and all data. This exhibits that the non-coding transcripts are equally important as the protein coding transcripts. Moreover, in recent studies, the dysregulation of long non-coding RNA in HCC has been studied [57] and their use as biomarkers has also been investigated [58].

Multiple machine learning algorithms (RF, NB, SVM, KNN, and NNET) were used to analyze the data and all exhibited high efficiency. It was surprising to see how well

these algorithms performed, despite significant variations in the sample and library preparation by different labs. Though all exhibited high efficiency, we observed some differences among them across all datasets as illustrated by various metrics calculated for them (Fig. 2 and 3). The reason for the varying performance of these algorithms on the same datasets can be explained by how their hyper-parameters are set. For instance, in the case of RF, the hyperparameters can be the number of samples required to split a node or tree depth; for KNN it can be the number of iterations to form k-groups or clusters; for NNET it can be node weights.

The highest values for all metrics were demonstrated by RF and SVM on all data-all transcripts dataset and the confidence intervals were smallest for RF for the mentioned dataset. NB also exhibited high values for all metrics for all data-all transcripts dataset however it performed poorly for other datasets and hence was not considered for further analyses. Hence top 32 important features were extracted from the algorithm-dataset combination (RF and SVM with all data-all transcripts) to find the least number of features using RFE. RFE employs a backward selection of the predictors, starting with all and removing the ones with the least importance in the model. Three transcripts were identified having maximum accuracy and kappa (Fig. 5B). None of these three transcripts were the ones that were taken randomly from correlated transcripts (c.f. Methodology 3b) and hence no transcript was discarded (correlation >0.75) that could have provided the same prediction accuracy. One of the chosen transcript was a non-coding transcript (CYREN-211). While many studies have emphasized the role of non-coding transcripts in the initiation, progression, and metastasis of HCC [59-62], their identification as key features to differentiate HCC and healthy liver is highlighted in only a handful of recent studies [63, 64].

Re-training the model using the three selected transcripts by applying exhaustive cross-validation helped in establishing their potency in discriminating the healthy from the HCC cell models. A final comparison with randomly selected highly expressed transcripts further established that these three transcripts were not chance findings; with values for all metrics always higher than the random combinations for RF and only 6 cases exhibited higher values for SVM. The three selected transcripts are involved in DNA repair pathways (CYREN-211 and PARP2-202) and immune response (SPON2-203). The DNA repair pathways are known to be affected in most cancers [65, 66] and in recent studies, immune dysfunction in

HCC and immunomodulation as a major factor in HCC development have been highlighted [67, 68].

Though these transcripts are validated through *in silico* approaches, an extensive validation in the HCC patients still needs to be done. If established, such an approach can also be used to identify transcript level biomarkers for various diseases and conditions, thus providing us an opportunity to look beyond proteins and maybe help in the identification of the disease or the condition at an early stage. One drawback of the current study was that the data was taken from the liver and to predict HCC, an invasive approach has to be taken to extract the sample. To look for transcript biomarkers for HCC that are non-invasive, data from HCC patient's blood serum/plasma will be required. At this moment, the scarcity of such data limits us from exploring the circulating mRNAs from HCC to find novel and potent biomarkers through *in silico* approaches. A thorough follow up study would be required to look for non-invasive/circulating transcript biomarkers in the blood of the HCC patients, by generating and analyzing the data as discussed in this study.

Conclusion

In our investigation of the healthy liver and various HCC cell models to find novel biomarkers, we analyzed RNA-Seq data using machine learning. Comparing the known HCC biomarkers with all other possible transcripts, we first concluded that using the exhaustive transcript list displayed better accuracy, thus implying that better biomarkers exist. Similarly, between all existing transcripts, protein coding transcripts only or non-coding transcripts only, it was illustrated that all transcriptomics data improved also the overall accuracy. From this observation, it can be concluded that both protein coding and non-coding transcripts hold important information and are regulated under internal and/or external stimuli. This is further supported by the identification of two protein coding (PARP2-202 and SPON2-203) and one non-coding (CYREN-211) transcript as novel and potent biomarker for HCC. However, the findings would have to be validated *in vivo*.

The pipeline developed in this study to identify transcript level biomarkers for HCC can be applied to other RNA-Seq datasets as well.

References

1. El-Serag HB: **Hepatocellular carcinoma**. *The New England journal of medicine* 2011, **365**(12):1118-1127.
2. Stewart B, Wild CP: **World cancer report 2014**. 2014.
3. Kochanek KD, Murphy SL, Xu J, Arias E: **Deaths: final data for 2017**. 2019.
4. Yang JD, Hainaut P, Gores GJ, Amadou A, Plymoth A, Roberts LR: **A global view of hepatocellular carcinoma: trends, risk, prevention and management**. *Nature Reviews Gastroenterology & Hepatology* 2019, **16**(10):589-604.
5. Roberts LR, Sirlin CB, Zaiem F, Almasri J, Prokop LJ, Heimbach JK, Murad MH, Mohammed K: **Imaging for the diagnosis of hepatocellular carcinoma: A systematic review and meta-analysis**. *Hepatology* 2018, **67**(1):401-421.
6. Bosman FT, Carneiro F, Hruban RH, Theise ND: **WHO classification of tumours of the digestive system**: World Health Organization; 2010.
7. Carr BI, Akkiz H, Üsküdar O, Yalçın K, Guerra V, Kuran S, Karaoğullarından Ü, Altıntaş E, Özakyol A, Tokmak S *et al*: **HCC with low- and normal-serum alpha-fetoprotein levels**. *Clin Pract (Lond)* 2018, **15**(1):453-464.
8. Wei W, Liu M, Ning S, Wei J, Zhong J, Li J, Cai Z, Zhang L: **Diagnostic value of plasma HSP90α levels for detection of hepatocellular carcinoma**. *BMC Cancer* 2020, **20**(1):6.
9. Zacharakis G, Aleid A, Aldossari KK: **New and old biomarkers of hepatocellular carcinoma**. *Hepatoma Res* 2018, **4**:65.
10. Toyoda H, Kumada T, Tada T, Sone Y, Kaneoka Y, Maeda A: **Tumor markers for hepatocellular carcinoma: simple and significant predictors of outcome in patients with HCC**. *Liver cancer* 2015, **4**(2):126-136.
11. Yu JP, Xu XG, Ma RJ, Qin SN, Wang CR, Wang XB, Li M, Li MS, Ma Q, Xu WW: **Development of a clinical chemiluminescent immunoassay for serum GPC3 and simultaneous measurements alone with AFP and CK19 in diagnosis of hepatocellular carcinoma**. *Journal of clinical laboratory analysis* 2015, **29**(2):85-93.
12. Tremosini S, Forner A, Boix L, Vilana R, Bianchi L, Reig M, Rimola J, Rodríguez-Lope C, Ayuso C, Solé M: **Prospective validation of an immunohistochemical panel (glypican 3, heat shock protein 70 and glutamine synthetase) in liver biopsies for diagnosis of very early hepatocellular carcinoma**. *Gut* 2012, **61**(10):1481-1487.
13. Mashaly AH, Anwar R, Ebrahim MA, Eissa LA, El Shishtawy MM: **Diagnostic and Prognostic Value of Talin-1 and Midkine as Tumor Markers in Hepatocellular Carcinoma in Egyptian Patients**. *Asian Pac J Cancer Prev* 2018, **19**(6):1503-1508.
14. Montagnana M, Danese E, Lippi G: **Squamous cell carcinoma antigen in hepatocellular carcinoma: Ready for the prime time?** *Clinica Chimica Acta* 2015, **445**:161-166.
15. Guo LY, Zhu P, Jin XP: **Association between the expression of HIF-1α and VEGF and prognostic implications in primary liver cancer**. *Genet Mol Res* 2016, **15**(2):15028107.
16. Zhou F, Shang W, Yu X, Tian J: **Glypican-3: A promising biomarker for hepatocellular carcinoma diagnosis and treatment**. *Medicinal research reviews* 2018, **38**(2):741-767.
17. Ai N, Liu W, Li ZG, Ji H, Li B, Yang G: **High expression of GP73 in primary hepatocellular carcinoma and its function in the assessment of transcatheter arterial chemoembolization**. *Oncology letters* 2017, **14**(4):3953-3958.
18. Lou J, Zhang L, Lv S, Zhang C, Jiang S: **Biomarkers for hepatocellular carcinoma**. *Biomarkers in cancer* 2017, **9**:1179299X16684640.
19. Wei R, Wong JPC, Kwok HF: **Osteopontin--a promising biomarker for cancer therapy**. *Journal of Cancer* 2017, **8**(12):2173.
20. Wang Q-s, Shi L-L, Sun F, Zhang Y-f, Chen R-W, Yang S-l, Hu J-l: **High Expression of ANXA2 Pseudogene ANXA2P2 Promotes an Aggressive Phenotype in Hepatocellular Carcinoma**. *Disease markers* 2019, **2019**.
21. Jin Y, Wang S, Chen W, Zhang J, Wang B, Guan H, Tang J: **Annexin A7 suppresses lymph node metastasis of hepatocarcinoma cells in a mouse model**. *BMC Cancer* 2013, **13**:522-522.
22. Mustika S, Wijaya H, Pratomo B: **The Expressions of CD44, CD90 and Alpha Fetoprotein Biomarkers in Indonesian Patients with Advanced Liver Disease: an Observational Study**. *Acta Medica Indonesiana* 2019, **51**(2):137-144.

23. Chen Y-L, Lin P-Y, Ming Y-Z, Huang W-C, Chen R-F, Chen P-M, Chu P-Y: **The effects of the location of cancer stem cell marker CD133 on the prognosis of hepatocellular carcinoma patients**. *BMC Cancer* 2017, **17**(1):474-474.
24. Noh C-K, Wang HJ, Kim CM, Kim J, Yoon SY, Lee GH, Cho HJ, Yang MJ, Kim SS, Hwang JC: **EpCAM as a Predictive Marker of Tumor Recurrence and Survival in Patients Who Underwent Surgical Resection for Hepatocellular Carcinoma**. *Anticancer research* 2018, **38**(7):4101-4109.
25. Chen J, Gingold JA, Su X: **Immunomodulatory TGF-β Signaling in Hepatocellular Carcinoma**. *Trends in molecular medicine* 2019.
26. Zheng N, Wei W, Wang Z: **Emerging roles of FGF signaling in hepatocellular carcinoma**. *Transl Cancer Res* 2016, **5**(1):1-6.
27. Kawaguchi M, Kataoka H: **Mechanisms of hepatocyte growth factor activation in cancer tissues**. *Cancers* 2014, **6**(4):1890-1904.
28. Li CN, Hsu HL, Wu TL, Tsao KC, Sun CF, Wu JT: **Cell-free DNA is released from tumor cells upon cell death: a study of tissue cultures of tumor cell lines**. *J Clin Lab Anal* 2003, **17**(4):103-107.
29. Yuan T, Huang X, Woodcock M, Du M, Dittmar R, Wang Y, Tsai S, Kohli M, Boardman L, Patel T *et al*: **Plasma extracellular RNA profiles in healthy and cancer patients**. *Sci Rep* 2016, **6**(19413).
30. Skog J, Würdinger T, van Rijn S, Meijer DH, Gainche L, Sena-Esteves M, Curry WT, Jr., Carter BS, Krichevsky AM, Breakefield XO: **Glioblastoma microvesicles transport RNA and proteins that promote tumour growth and provide diagnostic biomarkers**. *Nat Cell Biol* 2008, **10**(12):1470-1476.
31. Cheung KWE, Choi S-yR, Lee LTC, Lee NLE, Tsang HF, Cheng YT, Cho WCS, Wong EYL, Wong SCC: **The potential of circulating cell free RNA as a biomarker in cancer**. *Expert Review of Molecular Diagnostics* 2019, **19**(7):579-590.
32. Sayeed A, Dalvano BE, Kaplan DE, Viswanathan U, Kulp J, Janneh AH, Hwang L-Y, Ertel A, Doria C, Block T: **Profiling the circulating mRNA transcriptome in human liver disease**. *Oncotarget* 2020, **11**(23):2216-2232.
33. Read A, Natrajan R: **Splicing dysregulation as a driver of breast cancer**. *Endocr Relat Cancer* 2018, **25**(9):R467-R478.
34. Urbanski LM, Leclair N, Anczuków O: **Alternative-splicing defects in cancer: Splicing regulators and their downstream targets, guiding the way to novel cancer therapeutics**. *Wiley Interdiscip Rev RNA* 2018, **9**(4):e1476-e1476.
35. Jiménez-Vacas JM, Herrero-Aguayo V, Montero-Hidalgo AJ, Gómez-Gómez E, Fuentes-Fayos AC, León-González AJ, Sáez-Martínez P, Alors-Pérez E, Pedraza-Arévalo S, González-Serrano T: **Dysregulation of the splicing machinery is directly associated to aggressiveness of prostate cancer**. *EBioMedicine* 2020, **51**:102547.
36. Mjolsness E, DeCoste D: **Machine learning for science: state of the art and future prospects**. *Science* 2001, **293**(5537):2051-2055.
37. Kan A: **Machine learning applications in cell image analysis**. *Immunol Cell Biol* 2017, **95**(6):525-530.
38. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI: **Machine learning applications in cancer prognosis and prediction**. *Comput Struct Biotechnol J* 2014, **13**:8-17.
39. Zhang J, Naik HS, Assefa T, Sarkar S, Reddy RV, Singh A, Ganapathysubramanian B, Singh AK: **Computer vision and machine learning for robust phenotyping in genome-wide studies**. *Sci Rep* 2017, **7**(44048).
40. Ma C, Zhang HH, Wang X: **Machine learning for Big Data analytics in plants**. *Trends in plant science* 2014, **19**(12):798-808.
41. Ma C, Xin M, Feldmann KA, Wang X: **Machine learning-based differential network analysis: a study of stress-responsive transcriptomes in Arabidopsis**. *Plant Cell* 2014, **26**(2):520-537.
42. Zhang Z, Liu Z-P: **Identifying Cancer Biomarkers from High-Throughput RNA Sequencing Data by Machine Learning**. In: *Intelligent Computing Theories and Application: 2019// 2019; Cham*. Springer International Publishing: 517-528.
43. Wang L, Xi Y, Sung S, Qiao H: **RNA-seq assistant: machine learning based methods to identify more transcriptional regulated genes**. *BMC genomics* 2018, **19**(1):546.
44. Johnson NT, Dhroso A, Hughes KJ, Korkin D: **Biological classification with RNA-seq data: Can alternatively spliced transcript expression enhance machine learning classifiers?** *RNA (New York, NY)* 2018, **24**(9):1119-1132.
45. Akter S: **A Data Mining Approach for Biomarker Discovery Using Transcriptomics in Endometriosis**. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM): 2018*. IEEE: 969-972.
46. Bairoch A: **The cellosaurus, a cell-line knowledge resource**. *Journal of biomolecular techniques: JBT* 2018, **29**(2):25.

47. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data**. *Bioinformatics* 2014, **30**(15):2114-2120.
48. Frankish A, Vullo A, Zadissa A, Yates A, Thormann A, Parker A, Gall A, Moore B, Walts B, Aken BL *et al*: **Ensembl 2018**. *Nucleic Acids Res* 2017, **46**(D1):D754-D761.
49. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nat Methods* 2012, **9**(4):357-359.
50. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome**. *BMC Bioinformatics* 2011, **12**:323.
51. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**. *Genome Biol* 2014, **15**(12):550.
52. Kuhn M: **Caret: classification and regression training**. *ascl* 2015:ascl: 1505.1003.
53. John CR: **MLeval: Machine Learning Model Evaluation.** . 2020.
54. Jiao X, Sherman BT, Huang DW, Stephens R, Baseler MW, Lane HC, Lempicki RA: **DAVID-WS: a stateful web service to facilitate gene/protein list analysis**. *Bioinformatics* 2012, **28**(13):1805-1806.
55. Radley AH, Schwab RM, Tan Y, Kim J, Lo EKW, Cahan P: **Assessment of engineered cells using CellNet and RNA-seq**. *Nature Protocols* 2017, **12**:1089.
56. Chauhan R, Lahiri N: **Tissue- and Serum-Associated Biomarkers of Hepatocellular Carcinoma**. *Biomarkers in cancer* 2016, **8**(Suppl 1):37-55.
57. Huo X, Han S, Wu G, Latchoumanin O, Zhou G, Hebbard L, George J, Qiao L: **Dysregulated long noncoding RNAs (lncRNAs) in hepatocellular carcinoma: implications for tumorigenesis, disease progression, and liver cancer stem cells**. *Molecular Cancer* 2017, **16**(1):165.
58. Bao H, Su H: **Long Noncoding RNAs Act as Novel Biomarkers for Hepatocellular Carcinoma: Progress and Prospects**. *BioMed Research International* 2017, **2017**:6049480.
59. DiStefano JK: **Long noncoding RNAs in the initiation, progression, and metastasis of hepatocellular carcinoma**. *Noncoding RNA Res* 2017, **2**(3-4):129-136.
60. Wong C-M, Tsang FH-C, Ng IO-L: **Non-coding RNAs in hepatocellular carcinoma: molecular functions and pathological implications**. *Nature Reviews Gastroenterology & Hepatology* 2018, **15**(3):137-151.
61. Li C, Xu X: **Biological functions and clinical applications of exosomal non-coding RNAs in hepatocellular carcinoma**. *Cell Mol Life Sci* 2019, **76**(21):4203-4219.
62. He Y, Meng XM, Huang C, Wu BM, Zhang L, Lv XW, Li J: **Long noncoding RNAs: Novel insights into hepatocelluar carcinoma**. *Cancer Lett* 2014, **344**(1):20-27.
63. Li G, Shi H, Wang X, Wang B, Qu Q, Geng H, Sun H: **Identification of diagnostic long non-coding RNA biomarkers in patients with hepatocellular carcinoma**. *Molecular medicine reports* 2019, **20**(2):1121-1130.
64. Tan C, Cao J, Chen L, Xi X, Wang S, Zhu Y, Yang L, Ma L, Wang D, Yin J *et al*: **Noncoding RNAs Serve as Diagnosis and Prognosis Biomarkers for Hepatocellular Carcinoma**. *Clinical Chemistry* 2019, **65**(7):905-915.
65. Yang S-F, Chang C-W, Wei R-J, Shiue Y-L, Wang S-N, Yeh Y-T: **Involvement of DNA damage response pathways in hepatocellular carcinoma**. *BioMed Research International* 2014, **2014**:153867-153867.
66. Lin Z, Xu S-H, Wang H-Q, Cai Y-J, Ying L, Song M, Wang Y-Q, Du S-J, Shi K-Q, Zhou M-T: **Prognostic value of DNA repair based stratification of hepatocellular carcinoma**. *Sci Rep-Uk* 2016, **6**:25999-25999.
67. Hou J, Zhang H, Sun B, Karin M: **The immunobiology of hepatocellular carcinoma in humans and mice: Basic concepts and therapeutic implications**. *Journal of hepatology* 2020, **72**(1):167-182.
68. Roderburg C, Wree A, Demir M, Schmelzle M, Tacke F: **The role of the innate immune system in the development and treatment of hepatocellular carcinoma**. *Hepat Oncol* 2020, **7**(1):HEP17-HEP17.

Chapter 5

Expression and order of assembly of protein complexes – applying dynamic Bayesian networks to RNA-Seq data

Rajinder Gupta[1], Jos Kleinjans[1], Florian Caiment[1*]

1. Department of Toxicogenomics, School of Oncology and Developmental Biology (GROW), Maastricht University, Maastricht, The Netherlands

Under review: Proteins: Structure, Function, and Bioinformatics

Abstract

Protein complexes are the multi-molecule machinery of the biological system that are crucially involved in biological structures and functions. Moreover, the assembly of the complexes is an intricate process and is well regulated, however, it is less understood. Though functionally and structurally imperative to the biological system, they are not contemplated while assessing different biological systems and/or conditions. In this work through time-series RNA-Seq data, we first present the calculation of the protein complex expression and as an application of dynamic Bayesian networks, the prediction of assembly order of protein complexes. We show that the lowest expressed subunit at different time points can be the same or different subunit and it defines the expression of the complex at the given time point. For instance, the expression of SMAD2-SMAD3-SMAD4 complex (CPX-1) was shown to be determined bya different subunit of the complex during the time series. Furthermore, from the expression of individual subunits over different time points, dynamic Bayesian networks were generated and, analyzing the edges over consecutive time points, the order of assembly was predicted. For an arbitrarily chosen complex, Laminin211-nidogen complex (CPX-1282), analyzing the network suggested that the initial interaction happened between *LAMB1* and *NID1*, followed by *LAMC1* with *LAMA2*, and finally, *LAMB1* and *LAMC1* bind to each other. The information of protein complexes (expression and protein-complex assembly order) is imperative in assessing the biological function and identification of probable drug targets (conformational changes during assembly).

Introduction

Various proteins, macromolecules, metabolites, nucleic acids, and other biomolecules are produced within or transported inside the cells for its function and conformation structure. This crowded environment of the cell results in continuous interaction of all these biomolecules. Many interactions are non-specific [1, 2]; however, others form the basis for a variety of biological processes including, but not limited to, cell cycle regulation, differentiation, protein folding, translation, transcription, post-translational modifications, gene expression, enzyme inhibition, and antibody-antigen interactions [3, 4]. These bound molecules that possess important functions are termed protein complexes. The interactions between the subunits of a complex can be either short-lived, for instance in the case of molecules involved in signaling networks, or stable over time such as filaments. Moreover, the binding affinity, given by the dissociation constant (K_d), between the proteins varies and is influenced by various factors, such as cellular conditions, mutations in the protein, and protein folding. Several databases [5-11] and *in silico* [12-21] approaches are available to assess this binding affinity.

The complexes can also be classified as homomultimeric (homomeric) and heteromultimeric (heteromeric) protein complexes, according to their composition of the same or different biomolecules. Complexes can vary in their function and composition depending on the tissue type, developmental stage, and biological process, for instance, the lactate dehydrogenase (LDH) enzyme is a tetramer made of two different subunits, the H-form and the M-form. It assembles into five different complexes across various tissues in humans. LDH1 (4H; homomer) is found in the heart, LDH2 (3H1M; heteromer) in the reticuloendothelial system, LDH3 (2H2M; homomer) in the lungs, LDH4 (1H3M; heteromer) in the kidney, and LDH5 (4M; homomer) in the liver and striated muscle [22]. These five assemblies of the LDH differ in their electrophoretic mobility, with LDH1 being the fastest and LDH5 being the slowest.

While a complex can be homomeric or heteromeric, the number of units of the participating biomolecules can vary. This distribution of the number of biomolecules in a complex is known as stoichiometry and is an important aspect while studying the complexes. Most heteromeric complexes have an equal number of molecules of the subunits, known as even stoichiometry; however, a significant minority have uneven stoichiometry, implying differing numbers of each subunit

type. The DNA polymerase epsilon complex in *Saccharomyces cerevisiae* demonstrates even stoichiometry and has one chain of each DNA polymerase epsilon subunit A, B, C, and D. An example of the uneven stoichiometry is SMAD1-SMAD4 complex in humans, where SMAD1 has a stoichiometry of two and SMAD4 has one. Recent studies have demonstrated increased translational efficiency for the higher stoichiometry subunits within a complex [23, 24]. It has been shown that the most common uneven stoichiometry that exists are 2:1 (including 2:1, 4:2, 6:3, and similar) and 3:1 (includes 3:1, 6:2, 9:3, and similar) [25].

Using various approaches, namely, X-ray crystallography, nuclear magnetic resonance (NMR), various mass-spectrometry techniques such as native, cross-linked (CX or XL), ion mobility, and protein microarray [26]; the three-dimensional structures of thousands of protein complexes have been determined [27]. Additionally, there are also some computational methods available for protein complex prediction, these can be mainly divided into three categories: network-based, biological-context-aware, and specialized methods; reviewed extensively in [26]. This information on the three-dimensional structure and conformation of the complexes has a broad impact on our understanding of biological function and evolution [28]. In a recent paper on the organizing principles of the protein complexes, it has been shown that the assembly of the homomeric protein complexes can be classified into three basic types: dimerization, cyclization, fractional transition, and into two basic types for heteromeric complexes: heteromeric subunit addition, and non-stoichiometric transition [29]. The paper further develops all possible topologies of the homomeric and heteromeric protein complexes and develops a periodic table for them. Such information on the topologies of the complexes can help in understanding the evolution of the protein complexes. Moreover, it can help in predicting the quaternary structure of the complex hence formed from the given stoichiometry; highlighting potential constraints for multi-subunit docking and hybrid methods. Lastly, it can be helpful in bioengineering the complexes.

There are multiple databases available for the protein complexes [26] but many of them are not maintained anymore. The two major protein complex databases that are annotated manually and updated regularly are CORUM [30] and Complex portal [31]. They have 2417 (non-redundant) (CORUM) and 779 (Complex portal) protein complexes listed for humans, respectively (assessed on 14/08/202). While CORUM houses more complexes than Complex Portal, information of the stoichiometry of

the subunits is given as text in the description which makes it difficult to extract it. While this could be achieved with Natural Language Processing (NLP) tools, it is a time-consuming process. The information of the stoichiometry is important in assessing the expression of the complex. Complex Portal houses complexes for 25 organisms (assessed on 14/08/2020) and has stoichiometry defined for the listed protein complexes. These protein complexes include protein-only complexes as well as protein-small molecule (Chemical Entities of Biological Interest, ChEBI [32]) and protein-nucleic acid complexes. All complexes are derived from physical molecular interaction evidence extracted from the literature and cross-referenced in the entry, or by curator inference from information on homologs in closely related species or by inference from a scientific background. All complexes are tagged with Evidence and Conclusion Ontology codes to indicate the type of evidence available for each entry [31].

Various important aspects of the protein complexes such as half-life, binding affinity, stoichiometry, organizing principles, and conformations have already provided valuable information. However, the estimation of the expression and assembly order of the complexes would be required to further increase our understanding. The knowledge of the expression of the protein complex can be useful in establishing if the function of the complex is up- or downregulated. The expression-determining subunit for the protein complex is the lowest expressed subunit, after adjusting for the stoichiometry.

Furthermore, the knowledge of the assembly order of the protein complexes can provide us the opportunity to find novel binding sites for the drugs to control and treat various diseases. During the assembly of a protein complex, binding affinity and conformations change. The knowledge of these attributes can be beneficial. Moreover, the assembly order information can also help in understanding evolution by mapping gene fusion events [33]. Currently, using mass spectrometry techniques, the knowledge of subunits and their connectivity is illustrated [34]. However, due to the use of destabilizing solutions used for mass spectrometry, it is not an ideal system for assessing the assembly order. Moreover, at any given time, there will be multiple conformations of each complex and assessing one complex at a time is a time-consuming process. Through an in-silico approach, preliminary predictions can be made that can then be validated in lab.

We hypothesized that the expression of the subunits of a protein complex at multiple time points can provide information on the assembly of the complex

through the generation of probabilistic networks. To generate the probabilistic networks, dynamic Bayesian networks (DBN) were used on a time-series RNA-Seq dataset. A DBN is a Bayesian network (BN) that establishes the relationship/edges between the variables/nodes within the same and across different time steps. The edges are never directed from a later time point to an earlier time point, however, the forward prediction is possible. We used RNA-Seq data for this study instead of proteomics data because it has been shown that the protein complexes assemble co-translationally [35-37]. Additionally, the complexes are formed in a step-wise manner and the time-series data capture such information. The subunits that are required earlier will be formed first and then the other subunits will follow based on their requirement in the protein complex assembly. The RNA-Seq data were obtained from a baseline 3D human hepatic cell model (Primary Human Hepatocytes + Kuepfer cells Spheroids from InSphero®). The protein complex data (subunits and their respective stoichiometry) were taken from Complex Portal.

For this study, transcript expression is used instead of gene expression because of the heterogeneity of gene expression. However, since each gene can generate multiple transcripts, to choose the principal transcript (or isoform) from the given gene, we used the APPRIS annotation [38] (Box 1). APPRIS uses structural and functional features of the protein-coding genes together with cross-species conservation information to annotate the splice isoforms. However, when unavailable, the longest protein-coding transcript of the gene was selected as the principal transcript. From the expression of these transcripts, the subunit expression was calculated. The assembly order was then predicted using DBNs and analyzing the connections between the nodes over consecutive time points, the assembly order was predicted. The knowledge of the protein complex expression and the assembly order of the complex would add another dimension to the study of RNA-Seq data and will increase our understanding of the biological systems.

Methodology

Data

Protein complexes are taken from Complex Portal [31]. The total number of protein complexes obtained from the Complex Portal was 779 (as of 14/08/2020). Mainly, the protein complexes are made up of proteins; however, some may also include nucleic acids and/or small molecules. The Complex Portal gives the data for the

complexes as proteins defined by Uniprot identifiers, nucleic acids given as RNAcentral identifiers [39], and small molecules given as ChEBI identifiers [32].

We used transcriptomics data to calculate the expression of the protein complexes and their order of assembly. The RNA-Seq data were obtained from a baseline 3D human hepatic cell model (Primary Human Hepatocytes + Kuepfer cells Spheroids from InSphero®). Ribo-depleted libraries were generated from these cell models and sequenced on an Illumina Hiseq2000 in 100bp paired-end. Data are available in the BioStudies database (http://www.ebi.ac.uk/biostudies) under accession number S-HECA143. Eight-time points were analyzed: 0, 2, 8, 24, 72, 168, 240, and 336h and each time point had three replicates, making the total samples to 24. The quality assessment of the data has already been performed in one of our other study [40]. All samples passed the quality assessment. FPKM (Fragments Per Kilobase of transcript per Million mapped reads) was taken for calculation and prediction of the complex assembly order to normalize the expression for the length of the transcripts. All scripts and results are made available on https://github.com/rajinder4489/protein_complexes.

Filtration and mapping

The protein complexes were subjected to multiple steps of filtration (Fig. 1). First, the protein complexes having nucleic acid and/or small molecule as a subunit were discarded. In the transcriptomics data, small molecules (ChEBI) are not quantified; hence, the complexes that have small molecules as subunits (which are not quantified by a standard RNA sequencing libraries) are removed from further analyses. Similarly, complexes with non-coding RNAs were discarded because their functional annotation is limited.

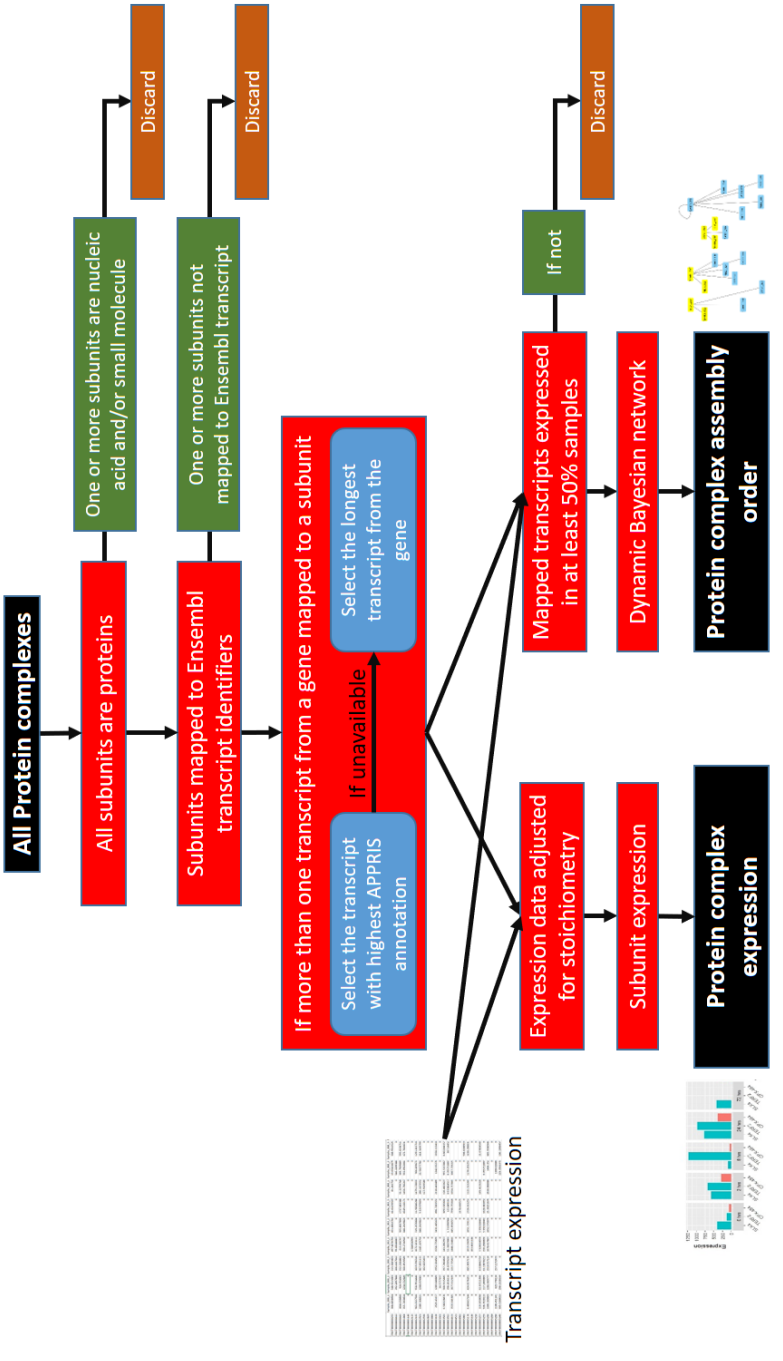


***Figure 1**: Workflow for calculation of protein complex expression and assembly order prediction.*

While the protein complexes from the Complex Portal are represented as protein subunits (along with other molecules) given by Uniprot identifiers [41], the RNA-Seq data is quantified in terms of Ensembl genes and transcripts identifiers [42]. To map these identifiers, Biomart [43] was used. When no transcript was mappable to a subunit, the protein complex was discarded. On the contrary, when more than one transcript mapped to the subunit, all mapped transcripts were retained for further processing. Additionally, there were cases where multiple transcripts from one gene mapped to the given subunit, which were addressed as described ahead.

Subunit expression

To resolve the cases of multiple transcripts from a gene mapping to one subunit, the transcript with the highest APPRIS annotation (Box 1) was selected. If more than one transcript from the same gene exhibit the same highest APPRIS annotation, all those transcripts were considered. Transcripts from different genes mapped to the same subunit and from all such genes, the highest APPRIS annotated transcript was taken. When no APPRIS annotation was available, the longest mapped transcript per gene was selected (or all the longest transcripts in case of several equally long transcript).

Protein complex expression

The protein complexes where at least one subunit was not detected in more than 50% of the samples, based on the RNA seq data, were discarded from further analysis. This step was performed to avoid making calculations using insufficient data. For the remaining protein complexes, the expression of each subunit was calculated per time point by averaging the expression of replicates for each time point. Furthermore, the expression was adjusted for the subunit stoichiometry by dividing the expression of the subunit by the corresponding stoichiometry. If the stoichiometry was unknown, it was set to one. Finally, the protein complex expression was set to the lowest expressed subunit per time point.

Order of assembly

Only heteromers should be taken for the prediction of the order of assembly because, in the case of homomers, only a single subunit is present and thus it cannot be used to define the order. DBNs were applied to the subunit expression before adjusting for the stoichiometry. The unadjusted subunit expression was used because the subunits with higher stoichiometry are produced at a higher amount to form the complex and the stoichiometry adjustment would reduce their impact in the DBN calculation. The R library bnstruct [44] was used to build the DBN, using the default parameters ('mmhc' algorithm and 'BDeu' (Bayesian-Dirichlet equivalent uniform) scoring function). The number of parents was limited to one and the edges between the same subunits over different time points were discarded. From the remaining network, the assembly order of the complexes was inferred by aligning the nodes/subunits over time using Cytoscape [45] and then analyzing the edges between the subunits over consecutive time points ($t \rightarrow t+1$). If the connection between all subunits could not be established by analyzing the consecutive time points, then the connections from $t \rightarrow t+2$ are analyzed and it is further expanded in a stepwise manner until the assembly order of all subunits could be predicted.

Results

Complex Portal houses 779 human protein complexes. The distribution of stoichiometry for these protein complexes was: 559 even, 122 uneven, and 98 undefined. For the complexes where this information was undefined, the stoichiometry was set to one for all subunits, labeling them as protein complexes with even stoichiometry. The protein complexes were subjected to various steps of filtration. First, the protein complexes comprising of only protein subunits were taken Then the protein complexes which had unmapped identifiers between Uniprot (protein complex subunits) and Ensembl (transcript identifiers) were removed from further analyses and finally the transcripts that were not expressed in at least 50% of the samples in the RNA-Seq dataset used in this study (Table 1) were discarded.

***Table 1**: Protein complexes through various steps of filtration and processing*

Step	**Number of protein complexes (discarded)**		**Action performed**
Downloaded from Complex Portal for *homo sapiens*	779		All protein complexes from Complex Portal downloaded
Complexes without nucleic acids and/or small molecules	681 (98)		-
Uniprot identifiers for all subunits mapped to at least one Ensembl transcript identifier	445 (236)		-
Selecting transcripts	445 (0)		Calculation of protein complex expression
	Transcripts for all subunits selected by APPRIS annotation	426	
	Transcripts for all subunits selected by Longest transcript	2	
	Transcripts for some subunits selected by APPRIS and for others by Longest transcript	17	
Homomers removed	384 (61)		-
Mapped transcripts detected in more than 50% of samples At least one mapped transcript per subunit is retained	69 (315)		Prediction of assembly order using bnstruct

Protein complex expression

After removing the complexes where the subunits were other than proteins and when a subunit could not be mapped to an Ensembl identifier, 445 protein complexes were obtained for calculation of the protein complex expression. The protein complex expression was determined by the lowest expressed subunit of the complex, after adjusting for the stoichiometry. In the case of the homomers (426), the expression of the complex was given by the expression of the subunit adjusted for the stoichiometry. As in the case of ANPR-A receptor complex (CPX-35) where

the stoichiometry of its subunit NPR1 was two, the complex expression was half of the expression of the lowest expressed subunit (figure not shown).

For 203 complexes out of 445 (45.62%), the complex expression was defined by a single subunit across all time points and for others (54.48%) by different subunits at different time points. Two examples are presented here, one where the stoichiometry of the subunits was one and the other where it was two, illustrating the impact of stoichiometry on the expression of the complex. First, for SMAD2-SMAD3-SMAD4 complex (CPX-1) all had a stoichiometry of 1 for all its subunits (SMAD2, SMAD3, and SMAD4) and the expression of the complex was defined by different lowest expressed subunits at each time point (Fig. 2A). SMAD2 was the lowest expressed subunit at time points 2, 8, 240, and 336 hours; SMAD3 at 168 hours, and SMAD4 for 0, 24, and 72 hours. Comparing the expression of the individual subunits exhibit higher perturbation than the protein complex and fail to present the functional perturbation. While SMAD2-SMAD3-SMAD4 complex had stoichiometry of one for all subunits, SLX4-TERF2 complex (CPX-484; subunits: SLX4 and TERF2) stoichiometry was two for both of its subunits (Fig. 2B). After adjusting for the stoichiometry, the expression of the complex was calculated and was given by SLX4 at time points 2, 8, 24, 168, and 336 hours and by TERF2 for 0, 72, and 240 hours. Out of 445 protein complexes, the complex was predicted to not expressed (expression = 0)for 185 complexes at all time-points because at least one of the subunits was not detected at the corresponding time points. Out of 185 complexes, 167 were heteromers and only 12 were homomers.

Furthermore, the expression profile of the complexes with respect to the subunits was evaluated using R library Genefilter [46]. The euclidean distance along with z-score scaling was used to calculate the distance between the expression profile of the subunits and the complex. For 252 (56.63%) complexes, the expression profile of the complex was similar to the expression profile of at least one subunit (Euclidean distance < 0.5) and for others, it was different from all subunits (193, 43.37%). The Laminin-521 complex (CPX-1780) has three subunits: LAMA5, LAMB2, and LAMC1; and the Euclidean distance between these three subunits and the complex was 0.49, 5.8, and 0.37, respectively. The smaller the Euclidean distance, the more is the similarity in the expression profile. As here LAMC1 has the closest expression profile to the Laminin-521 complex (Fig. 3A).

***Figure 2**: Expression of the subunits and the complexes. (A) For the protein complex SMAD2-SMAD3-SMAD4 (CPX-1), the stoichiometry was one for all subunits. The expression of the complex was determined by SMAD2 for 2, 8, 240, and 336 hours, by SMAD3 for 168 hours, and by SMAD4 for 0, 24, and 72 hours. (B) In the case of the SLX4-TERF2 complex (CPX-484), the stoichiometry of both the subunits was two and hence the expression of the subunits was adjusted for calculating the expression of the complex. The expression of the complex was given by SLX4 for 2, 8, 24, 168, and 336 hours and by TERF2 for 0, 72, and 240 hours.*

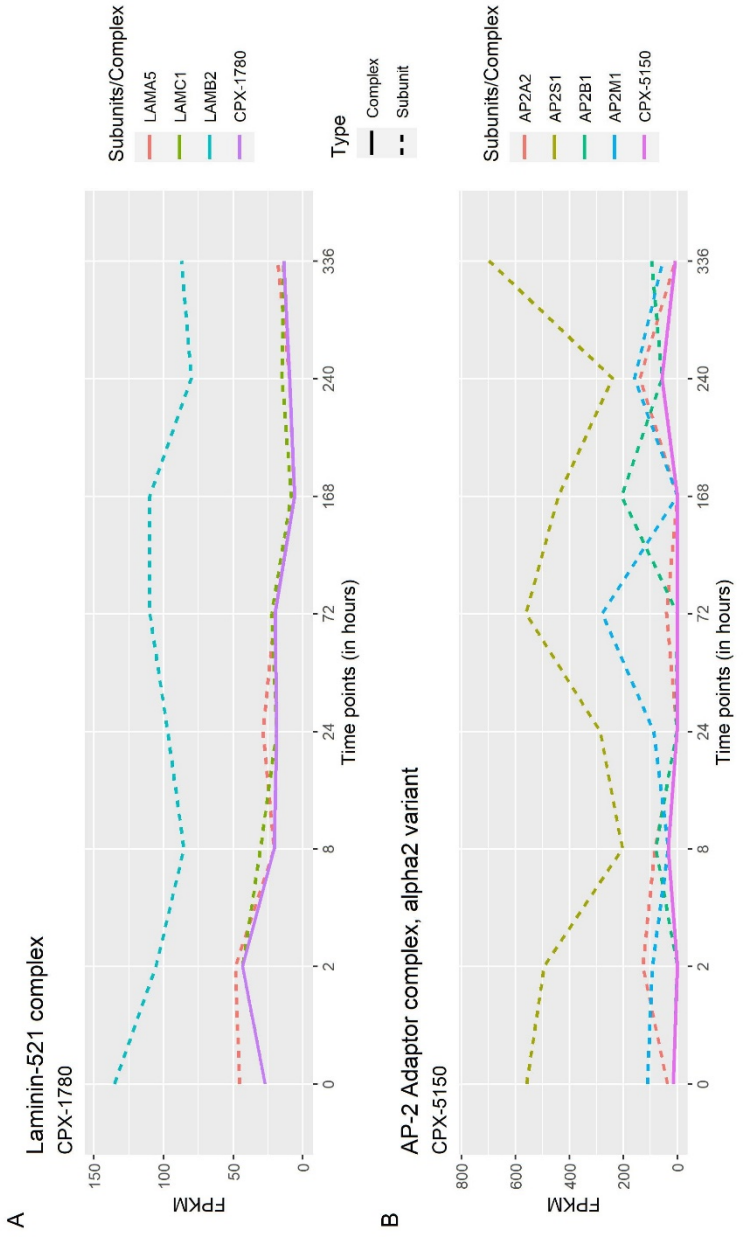


***Figure 3**: Expression profile over time. (A) For Laminin-521 complex (CPX-1780), the expression profile of the complex is similar to two subunits: LAMC1 and LAMA5. The third subunit LAMB2 presents a different expression profile. (B) In the case of AP-2 Adaptor complex, alpha2 variant (CPX-5150), the complex's expression profile is different from all subunits.*

On the other hand, AP-2 Adaptor complex, alpha2 variant (CPX-5150), there was no similarity between the expression profile of the complex and the subunits (Fig. 3B). The Euclidean distances for the four subunits to the complex were AP2A2: 1.44, AP2B1: 1.28, AP2M1: 1.98, and AP2S1: 6.64. The Euclidean distance for all subunits to their respective complexes is given in Suppl. Table 1.

Order of assembly

The protein complexes are formed by binding of the subunits in a phased manner. The subunits binding first make suitable conformational changes for the next subunit to bind to the nascent complex [47-49]. We predicted the assembly order using the DBNs. Only heteromers for which the transcripts were expressed in at least 50% of samples were taken, resulting in 69 complexes that were taken for assembly order prediction. Among the 69 protein complexes, only three complexes, namely SMAD1-SMAD4 complex (CPX-54), Collagen type I trimer (CPX-1650), and PDGF receptor beta - PDGF-AB complex (CPX-2886) had uneven stoichiometry of 2:1, 2:1, and 2:1:1 respectively. For these 69 complexes, DBNs were computed using bnstruct [44] and adjacency matrices are provided for all of them in Suppl. Table 2. The edges between the same subunits over different time points should be removed while predicting the order of assembly and the remaining network should be analyzed to assess earlier and later interactions.

In the case of an arbitrarily chosen complex - Laminin211-nidogen complex (CPX-1282) (Fig. 4), the protein complex consisted of four subunits: LAMA2, LAMB1, LAMC1, and NID1. The analyses of the DBN for consecutive time points ($t \rightarrow t+1$) predicted that LAMB1 (0h) interacts with NID1 (2h) and LAMC1 (2h) interacts with LAMA2 (2h and 8h). However, since no interactions were predicted between LAMB1-LAMA2, LAMB1-LAMC1, NID1-LAMA2, or NID1-LAMC1, the analyses of the DBN was expanded to $t \rightarrow t+2$. It was then predicted that LAMB1 (0h) interacts with LAMC1 (8h). Hence, DBN predicts that the overall assembly of this complex could be the initial formation of two nascent complexes: LAMB1-NID1, and LAMC1-LAMA2, followed by LAMB1 and LAMC1 binding to each other. This binding between LAMB1 and LAMC1 can be between the nascent complexes formed.

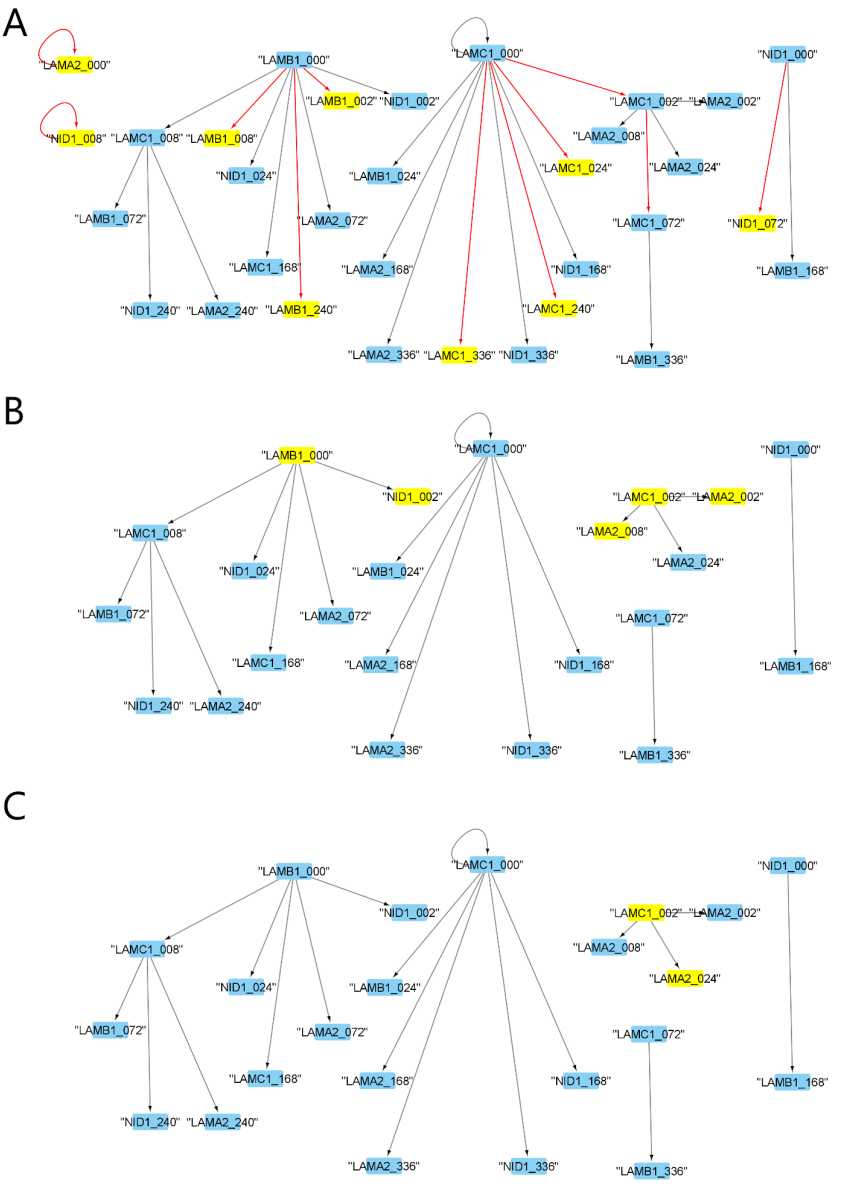


***Figure 4**: Protein-complex expression order prediction for a multimeric heteromer. The protein complex Laminin211-nidogen complex (CPX-1282) constitutes of four subunits: LAMA2, LAMB1, LAMC1, and NID1. (A) The DBN computed from the expression data for the subunits using R library bnstruct. The highlighted nodes and edges were deleted because they connect the same subunits over different time points (B) The order of assembly was predicted by analyzing the edges/connections between the nodes over consecutive time points (t → t+1). The nodes over initial consecutive time points are highlighted. They show that LAMC1 interacts with LAMA2 and LAMB1interacts with NID1. (C) The time points analyzed were expanded further (t → t+2). The highlighted nodes show that LAMB1 interacts with LAMC1.*

Discussion

Protein complexes are the functional and structural machinery of the biological system. The knowledge of their expression can be advantageous in studying biological functions and structures. Through this work, we presented an approach to calculate the expression of the protein complexes using the RNA-Seq time-series data. To circumvent the heterogeneity in the gene expression calculation, we used transcript expression for the calculation of expression of the protein complexes. Furthermore, as hypothesized, the probabilistic models generated from time-series expression data can help in predicting the assembly order of the protein complexes. We demonstrated a DBN-based approach to infer the assembly order of the protein complexes from time-series RNA-Seq data (eight time-points). The protein complexes are formed in a well-ordered manner and each binding subunit causes conformational changes in the nascent complex [47-49]. This creates favorable conditions for the next subunit to bind and helps the complex in attaining its structure and function.

We demonstrated that for ~54% of the complexes, the complex expression determining subunit is different at different time points. As illustrated for SMAD2-SMAD3-SMAD4 complex (CPX-1) and SLX4-TERF2 complex (CPX-484), the expression of the complexes were defined by different subunits over different time points. While the individual subunits exhibited higher rates of perturbation across various time points, the comparison of the complex expression helped in eliminating the false inferences derived from the subunit expression comparison. It has been shown that the complexes are formed co-translationally [35-37] and hence the higher changes in the expression of the subunits can be due to their demand in the complex assembly.

The expression of the complex was different from the subunits in various instances however, the expression profiles over time were similar as shown in Fig. 3A for Laminin-521 complex. In such cases, the analyses of the subunits can also be used to infer the functional properties of the biological system. The significance of complex expression analyses arises when the subunit and the complex expression profiles are different as presented in Fig. 3B for AP-2 Adaptor complex, alpha2 variant. The changes in the expression of the subunit might not be related to its function, rather to its demand in complex assembly.

The protein complex expression computed from the RNA-Seq cannot give the absolute expression of the protein complex in the biological system. However, it can be used as a comparative metric for estimating the functional changes. Multiple processes and steps of regulation are involved in the journey of a protein-coding transcript to form a protein. Moreover, there are also other limiting factors such as the amount of ribosomes, the half-life of transcripts and proteins, translational efficiency, and Spatio-temporal availability. The protein half-life of the different subunits could be different [50, 51]. The amount and saturation of the mRNA by ribosomes can indeed be a limiting factor in the amount of protein expressed [52]. In a study evaluating the half-lives of 803 proteins, it was shown that the median half-life of the proteins was 8.7 hours [53]. Moreover, the translational efficiency for different transcripts might be variable [54-56].

Another important aspect of the protein complexes is their order of assembly. It was computed using DBNs because they allow only the nodes from the earlier time points to be the parent of nodes in the same or later time points. No node from a later point can be a parent to a node in the earlier time point; hence the relationship moves forward in time. The analysis of the DBNs by studying the edges helped in establishing earlier and later connections between the subunits. The distribution of time points of RNA-Seq time-series data is based on PBPK modeling, however, RNA-Seq data collected at smaller and equal intervals may better exhibit the assembly of the protein complex. The correct prediction of the protein complex assembly might help in finding novel drug targets, as the binding of the subunits to the nascent protein complex results in conformational changes. These conformational changes occur to allow the presentation of the binding sites for incoming subunits and exposing the functional domains of the complex. Knowledge of what subunits bind and when they bind may help in assessing the conformational changes and hence their role as drug targets.

The current approach can be improved by identifying the nascent protein complex being formed as an individual entity in the computation of the network and assembly order prediction. Currently, only subunits are considered for the creation of DBNs. The expression of the forming complex can be given by the bound subunits at the given time point. Other options to explore the protein complex assembly order can be using the Bayesian networks for individual time points and then analyzing the high probability edges over each timepoint individually. This will be different from the DBN because there will not be any connections between the

nodes over different time points. Additionally, other temporal network prediction approaches can be used; some packages available in R to perform temporal analysis are sna [57], ndtv [58], and tsna [59]. To further enhance the creation of the networks, additional information can be added such as protein-protein interactions, ontology, and known functions.

In addition to advancing the computational approaches to predict the assembly order of the protein complexes, approaches to extract nascent protein complexes and investigating them through X-ray crystallography and mass spectrometry experiments to validate the *in silico* findings are equally important. Such a validation step would provide proof of the hypothesis and help in defining the order of assembly in other complexes.

Conclusion

The knowledge of the protein complex expression and their assembly order can pave the way for better quantification of the biological systems' functional capability and capturing the perturbations in protein complex assembly. Moreover, the conformational changes in the forming complex can be studied for their use as novel drug targets. However, further lab investigation of these in-silico findings is pertinent.

Box 1

APPRIS categories for the transcripts, taken/adapted from https://www.ensembl.org/Help/Glossary?id=521

APPRIS category (sorted by confidence; high to low)	Explanation
Principal1	Transcript(s) expected to code for the main functional isoform based solely on the core modules in the APPRIS.
Principal2	Where the APPRIS core modules are unable to choose a clear principal variant (approximately 25% of human protein coding

	genes), the database chooses two or more of the CDS variants as "candidates" to be the principal variant.
Principal3	Where the APPRIS core modules are unable to choose a clear principal variant and there more than one of the variants have distinct CCDS identifiers, APPRIS selects the variant with lowest CCDS identifier as the principal variant. The lower the CCDS identifier, the earlier it was annotated.
Principal4	Where the APPRIS core modules are unable to choose a clear principal CDS and there is more than one variant with a distinct (but consecutive) CCDS identifiers, APPRIS selects the longest CCDS isoform as the principal variant.
Principal5	Where the APPRIS core modules are unable to choose a clear principal variant and none of the candidate variants are annotated by CCDS, APPRIS selects the longest of the candidate isoforms as the principal variant.
Alternative1	For genes in which the APPRIS core modules are unable to choose a clear principal isoform, the alternative1 is the candidate transcript(s) models that is conserved in at least three tested species.
Alternative2	For genes in which the APPRIS core modules are unable to choose a clear principal isoform, the alternative2 is the candidate transcript(s) models that appear to be conserved in fewer than three tested species.

References

1. Johansson H, Jensen MR, Gesmar H, Meier S, Vinther JM, Keeler C, Hodsdon ME, Led JJ: **Specific and nonspecific interactions in ultraweak protein-protein associations revealed by solvent paramagnetic relaxation enhancements**. *J Am Chem Soc* 2014, **136**(29):10277-10286.
2. Johnson ME, Hummer G: **Nonspecific binding limits the number of proteins in a cell and shapes their interaction networks**. *Proceedings of the National Academy of Sciences* 2011, **108**(2):603-608.
3. Srihari S, Leong HW: **A survey of computational methods for protein complex prediction from protein interaction networks**. *Journal of bioinformatics and computational biology* 2013, **11**(02):1230002.
4. Ramyachitra D, Banupriya D: **Protein Complex Detection: A Study**. *International Journal of Computer Science and Information Technology & Security (IJCSITS)* 2014, **4**(4).
5. Thorn KS, Bogan AA: **ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions**. *Bioinformatics* 2001, **17**(3):284-285.
6. Kumar MS, Gromiha MM: **PINT: protein–protein interactions thermodynamic database**. *Nucleic Acids Res* 2006, **34**(suppl_1):D195-D198.
7. Moal IH, Fernández-Recio J: **SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models**. *Bioinformatics* 2012, **28**(20):2600-2607.
8. Vreven T, Moal IH, Vangone A, Pierce BG, Kastritis PL, Torchala M, Chaleil R, Jiménez-García B, Bates PA, Fernandez-Recio J: **Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2**. *Journal of molecular biology* 2015, **427**(19):3031-3041.

9. Liu Z, Li Y, Han L, Li J, Liu J, Zhao Z, Nie W, Liu Y, Wang R: **PDB-wide collection of binding data: current status of the PDBbind database**. *Bioinformatics* 2015, **31**(3):405-412.
10. Jankauskaitė J, Jiménez-García B, Dapkūnas J, Fernández-Recio J, Moal IH: **SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation**. *Bioinformatics* 2019, **35**(3):462-469.
11. Jemimah S, Yugandhar K, Michael Gromiha M: **PROXiMATE: a database of mutant protein–protein complex thermodynamics and kinetics**. *Bioinformatics* 2017, **33**(17):2787-2788.
12. Moal IH, Agius R, Bates PA: **Protein–protein binding affinity prediction on a diverse set of structures**. *Bioinformatics* 2011, **27**(21):3002-3009.
13. Moal IH, Jiménez-García B, Fernández-Recio J: **CCharPPI web server: computational characterization of protein–protein interactions from structure**. *Bioinformatics* 2015, **31**(1):123-125.
14. Yugandhar K, Gromiha MM: **Protein–protein binding affinity prediction from amino acid sequence**. *Bioinformatics* 2014, **30**(24):3583-3589.
15. Dehouck Y, Kwasigroch JM, Rooman M, Gilis D: **BeAtMuSiC: prediction of changes in protein–protein binding affinity on mutations**. *Nucleic Acids Res* 2013, **41**(W1):W333-W339.
16. Berliner N, Teyra J, Çolak R, Lopez SG, Kim PM: **Combining structural modeling with ensemble machine learning to accurately predict protein fold stability and binding affinity effects upon mutation**. *PloS one* 2014, **9**(9):e107353.
17. Brender JR, Zhang Y: **Predicting the effect of mutations on protein-protein binding interactions through structure-based interface profiles**. *PLoS Comput Biol* 2015, **11**(10):e1004494.
18. Petukh M, Dai L, Alexov E: **SAAMBE: webserver to predict the charge of binding free energy caused by amino acids mutations**. *International journal of molecular sciences* 2016, **17**(4):547.
19. Li M, Simonetti FL, Goncearenco A, Panchenko AR: **MutaBind estimates and interprets the effects of sequence variants on protein–protein interactions**. *Nucleic Acids Res* 2016, **44**(W1):W494-W501.
20. Pires DE, Ascher DB: **mCSM-AB: a web server for predicting antibody–antigen affinity changes upon mutation with graph-based signatures**. *Nucleic Acids Res* 2016, **44**(W1):W469-W473.
21. Rodrigues CHM, Myung Y, Pires DEV, Ascher DB: **mCSM-PPI2: predicting the effects of mutations on protein–protein interactions**. *Nucleic Acids Res* 2019, **47**(W1):W338-W344.
22. Dzoyem JP, Kuete V, Eloff JN: **23 - Biochemical Parameters in Toxicological Studies in Africa: Significance, Principle of Methods, Data Interpretation, and Use in Plant Screenings**. In: *Toxicological Survey of African Medicinal Plants.* Edited by Kuete V: Elsevier; 2014: 659-715.
23. Quax TE, Wolf YI, Koehorst JJ, Wurtzel O, van der Oost R, Ran W, Blombach F, Makarova KS, Brouns SJ, Forster AC *et al*: **Differential translation tunes uneven production of operon-encoded proteins**. *Cell Rep* 2013, **4**(5):938-944.
24. Li G-W, Burkhardt D, Gross C, Weissman JS: **Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources**. *Cell* 2014, **157**(3):624-635.
25. Marsh JA, Rees HA, Ahnert SE, Teichmann SA: **Structural and evolutionary versatility in protein complexes with uneven stoichiometry**. *Nature communications* 2015, **6**(1):1-10.
26. Zahiri J, Emamjomeh A, Bagheri S, Ivazeh A, Mahdevar G, Tehrani HS, Mirzaie M, Fakheri BA, Mohammad-Noori M: **Protein complex prediction: A survey**. *Genomics* 2019.
27. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic Acids Res* 2000, **28**(1):235-242.
28. Marsh JA, Teichmann SA: **Structure, dynamics, assembly, and evolution of protein complexes**. *Annu Rev Biochem* 2015, **84**:551-575.
29. Ahnert SE, Marsh JA, Hernández H, Robinson CV, Teichmann SA: **Principles of assembly reveal a periodic table of protein complexes**. *Science* 2015, **350**(6266).
30. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes H-W: **CORUM: the comprehensive resource of mammalian protein complexes—2009**. *Nucleic Acids Res* 2009, **38**(suppl_1):D497-D501.
31. Meldal BHM, Forner-Martinez O, Costanzo MC, Dana J, Demeter J, Dumousseau M, Dwight SS, Gaulton A, Licata L, Melidoni AN *et al*: **The complex portal - an encyclopaedia of macromolecular complexes**. *Nucleic Acids Res* 2014, **43**(D1):D479-D484.
32. Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C: **ChEBI in 2016: Improved services and an expanding collection of metabolites**. *Nucleic Acids Res* 2015, **44**(D1):D1214-D1219.
33. Marsh JA, Hernández H, Hall Z, Ahnert SE, Perica T, Robinson CV, Teichmann SA: **Protein complexes are under evolutionary selection to assemble via ordered pathways**. *Cell* 2013, **153**(2):461-470.

34. Gingras AC, Gstaiger M, Raught B, Aebersold R: **Analysis of protein complexes using mass spectrometry**. *Nat Rev Mol Cell Biol* 2007, **8**(8):645-654.
35. Mayr C: **Protein complexes assemble as they are being made**. In.: Nature Publishing Group; 2018.
36. Williams NK, Dichtl B: **Co-translational control of protein complex formation: a fundamental pathway of cellular organization?** *Biochemical Society Transactions* 2018, **46**(1):197-206.
37. Wells JN, Bergendahl LT, Marsh JA: **Co-translational assembly of protein complexes**. *Biochemical Society Transactions* 2015, **43**(6):1221-1226.
38. Rodriguez JM, Rodriguez-Rivas J, Di Domenico T, Vázquez J, Valencia A, Tress ML: **APPRIS 2017: principal isoforms for multiple gene sets**. *Nucleic Acids Res* 2017, **46**(D1):D213-D217.
39. **RNAcentral: a comprehensive database of non-coding RNA sequences**. *Nucleic Acids Res* 2016, **45**(D1):D128-D134.
40. Gupta R, Schrooders Y, Verheijen M, Roth A, Kleinjans J, Caiment F: **FuSe: A tool to move RNA-Seq analyses from chromosomal/gene loci to functional grouping of mRNA transcripts**. *Bioinformatics* 2020.
41. The_UniProt_Consortium: **UniProt: a hub for protein information**. *Nucleic Acids Res* 2015, **43**(Database issue):27.
42. Frankish A, Vullo A, Zadissa A, Yates A, Thormann A, Parker A, Gall A, Moore B, Walts B, Aken BL *et al*: **Ensembl 2018**. *Nucleic Acids Res* 2017, **46**(D1):D754-D761.
43. Smedley D, Haider S, Durinck S, Pandini L, Provero P, Allen J, Arnaiz O, Awedh MH, Baldock R, Barbiera G *et al*: **The BioMart community portal: an innovative alternative to large, centralized data repositories**. *Nucleic Acids Res* 2015, **43**(W1):20.
44. Franzin A, Sambo F, Di Camillo B: **bnstruct: an R package for Bayesian Network structure learning in the presence of missing data**. *Bioinformatics* 2017, **33**(8):1250-1252.
45. Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T: **Cytoscape 2.8: new features for data integration and network visualization**. *Bioinformatics* 2011, **27**(3):431-432.
46. Gentleman R, Carey V, Huber W, Hahne F, Maintainer MBP, AnnotationDbi I: **Package 'genefilter'**. 2013.
47. Ito Y, Ikeguchi M: **Mechanism of the αβ conformational change in F1-ATPase after ATP hydrolysis: free-energy simulations**. *Biophys J* 2015, **108**(1):85-97.
48. Chu W-T, Chu X, Wang J: **Binding mechanism and dynamic conformational change of C subunit of PKA with different pathways**. *Proceedings of the National Academy of Sciences* 2017, **114**(38):E7959-E7968.
49. Garton M, MacKinnon SS, Malevanets A, Wodak SJ: **Interplay of self-association and conformational flexibility in regulating protein function**. *Philosophical Transactions of the Royal Society B: Biological Sciences* 2018, **373**(1749):20170190.
50. Yang E, van Nimwegen E, Zavolan M, Rajewsky N, Schroeder M, Magnasco M, Darnell JE, Jr.: **Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes**. *Genome Res* 2003, **13**(8):1863-1872.
51. Sharova LV, Sharov AA, Nedorezov T, Piao Y, Shaik N, Ko MSH: **Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells**. *DNA Res* 2009, **16**(1):45-58.
52. Lin J, Amir A: **Homeostasis of protein and mRNA concentrations in growing cells**. *Nature communications* 2018, **9**(1):4496-4496.
53. Chen W, Smeekens JM, Wu R: **Systematic study of the dynamics and half-lives of newly synthesized proteins in human cells**. *Chemical science* 2016, **7**(2):1393-1400.
54. Mead EJ, Smales CM: **1.29 - mRNA Translation and Recombinant Gene Expression from Mammalian Cell Expression Systems**. In: *Comprehensive Biotechnology (Second Edition).* Edited by Moo-Young M. Burlington: Academic Press; 2011: 403-409.
55. Neves D, Vos S, Blank LM, Ebert BE: **Pseudomonas mRNA 2.0: Boosting Gene Expression Through Enhanced mRNA Stability and Translational Efficiency**. *Frontiers in Bioengineering and Biotechnology* 2020, **7**(458).
56. Cenik C, Cenik ES, Byeon GW, Grubert F, Candille SI, Spacek D, Alsallakh B, Tilgner H, Araya CL, Tang H *et al*: **Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans**. *Genome Res* 2015, **25**(11):1610-1621.
57. Butts CT: **Social network analysis with sna**. *Journal of statistical software* 2008, **24**(6):1-51.
58. Bender-deMoll S: **Package Vignette for ndtv: Network Dynamic Temporal Visualizations (Version 0.13. 0)**. 2019.
59. Skye Bender-deMoll MM, James Moody: **tsna: Tools for Temporal Social Network Analysis**. *R package version* 2020.

Chapter 6

Discussion

Gene expression is omnipresent with transcriptomics data analysis. However, gene expression data is heterogeneous, as it is obtained by summing the expression of all the transcripts it codes. Through this thesis, a transcript-based analysis of the RNA sequencing (RNA-Seq) data is presented and novel methods for data analysis and their applications are proposed. To test the novel methodologies developed in this thesis, the liver was chosen because of its significance in the human body and its predominance in the toxicology field. RNA-Seq data from various *in vitro* liver cell models generated under the aegis of EU-ToxRisk project and available through public databases formed the basis of this thesis.

In vitro cell models – the dilemma

Multiple challenges and limitations with animal testing for the drug efficacy and toxicity studies have highlighted the need for finding better testing models [1, 2]. The *in vitro* cell models have merged as a viable alternative to animal testing. A major problem with the *in vitro* cell models is their abundance. Various cell models for different tissues, cells, and conditions (such as cancer, knockdown) have been developed. The problem of plenty poses a challenge in selecting the best cell model for a study or research question. The selection is primarily based on prior knowledge and literature review, however, the literature is biased towards a well-studied set of genes, proteins, or processes. There is a gap in the knowledge to assess the cell models.

To be able to make an informed decision while selecting the cell models for a study, a thorough investigation of the *in vitro* liver cell models using the RNA-Seq data was performed in **Chapter 2**. Various *in vitro* cell models from the liver ranging from cancer derived (HepG2, HepaRG, hPCLiS (human precision cut liver slices)), iPSC derived, microtissues, PHH, and healthy *in vivo* liver were taken for this study. The healthy *in vivo* liver samples were used as control.

A global comparison was performed using correlation and CellNet [3], highlighting the similarity of the transcriptome profile to *in vivo* human liver through various metrics. First, through correlation, it was established that all cell models exhibited high similarity among themselves and to the *in vivo* liver. There was an even higher similarity between the replicates from the cell models. Overall, all cell models presented a spearman's correlation greater than 0.8 (range 0-1). Furthermore, using CellNet, various aspects of the cell models were compared – the similarity of

the overall transcriptome profile defined by C/T (Cell/Tissue) score, the conservation of the gene regulatory networks (GRN) given by GRN score, and the network influence score (NIS) of the transcription factors (TFs). The C/T score presents how similar the given cell model's RNA-Seq data is to various cells or tissues in the CellNet's consensus profile. As expected, the samples directly derived from the liver (healthy liver (PHH, 3D liver microtissues), and cancerous liver (hPCLiS)) illustrated a high C/T score while cancer- (HepG2, HepaRG 3D) and iPSC-derived (iPSC-HLC) models had a comparatively lower score with the liver; highlighting their extent of similarity with liver (c.f. Chapter 2 Fig. 2).

Similarly, the GRN score was used for comparing the cell models for the conservation of gene regulatory networks (c.f. Chapter 2 Suppl. Fig. 3) and NIS score evaluated the influence of perturbed TFs on the cell models (c.f. Chapter 2 Suppl. Fig. 5a-g). The identification of the perturbed TFs provides an opportunity to regulate their expression to make the cell models similar to the target cell/tissue type. For all cell models, the most perturbed TF was ATF5; moreover, in all cell models except PHH, it was shown that ATF5 was downregulated. With the current data, it was not possible to investigate the reason why ATF5 was perturbed in opposite direction as compared to other cell models. However, investigation of the culture conditions and/or protocols may elucidate the cause. Anyhow, the identification of the TFs that are perturbed and their degree of perturbation helps in developing the cell models to look and behave similarly to the target cell/tissue. Especially, in the case of iPSC cell models, the identification of the TFs that have different expression patterns as compared to target cell/tissue helps in adjusting their differentiation protocols and hence improving further their similarity to *in vivo* liver.

Furthermore, a comparison on the gene-level was achieved by analyzing differentially expressed genes (DEGs) and non-DEGs. While a higher number of DEGs corresponds to high dissimilarity, more non-DEGs demonstrate the similarity between the two cell models. The number of DEGs (most: HepG2, ~9900; least: 3D liver microtissues 0h, ~5200) and non-DEGs (most: 3D liver microtissues 0h, ~10200; least: HepG2, ~5700) only provides an overview of resemblance and fails to convey what processes are affected. To investigate the changes in the processes and pathways, the DEGs and non-DEGs were mapped to the KEGG human pathways and the extent of coverage illustrated whether the particular pathway was perturbed or not. On a total of ~420 human KEGG pathways [4], the list of genes

(DEGs and non-DEGs) from each cell model was mapped. Through this exercise, the pathway coverage for all human KEGG pathways was achieved and illustrated the extent of similarity/perturbation in each pathway for each cell model. Among the important liver pathways (c.f. Fig. 5 and Fig. 7, Chapter 2), the highest similarity was predominantly exhibited by PHH and 3D liver microtissues. The DNA-repair pathways were highly conserved in HepG2, HepaRG 3D, and hPCLiS. Surprisingly, iPSC-HLC demonstrated higher similarity than HepG2 and occasionally better than HepaRG 3D and hPCLiS. These analyses will help in selecting the appropriate cell models for studying a process and/or set of genes.

Beyond the genes – transcript expression

For transcriptomics data analyses, the convention has been to assess the gene expression and study its perturbation across conditions or samples. However, the heterogeneity in the gene expression, due to multiple transcripts originating from a gene, makes it a less reliable metric to assess. We explored the opportunities in a transcript-based analysis. Usually, the transcripts are studied for their differential expression (DET: differentially expressed transcripts, similar to DEGs) or usage (DTU: differential transcript usage). In DET analyses, the change in the amount of expression of the transcript between two conditions is compared while in DTU the change in the fraction of the transcript expression in the gene expression is calculated. Irrespective of DET or DTU analyses, the resulting transcripts cannot be mapped to biological databases (pathway, ontology, or function) as the databases are gene- and/or protein-based. To counter this challenge, we derived non-DEGs^{DTU-} (non-differentially expressed genes without differential transcript usage) by taking the non-DEGs and then removing the genes for which transcripts exhibited differential transcript usage. The resulting list of genes (non-DEGs^{DTU-}) comprised of the genes that were not differentially expressed on the gene and transcript level. Non-DEGs^{DTU-} could then be mapped to pathways and processes across various biological databases. In comparison to the pathway coverage by non-DEGs, non-DEG^{DTU-} exhibited a lower coverage for all pathways, however, it followed a similar trend as non-DEGs across all cell models.

Analyses of the transcript usage highlighted that, for many genes, the main protein-coding transcript from the gene is replaced by another protein-coding transcript or by a non-coding transcript. For instance, in the case of *POLR2F*, which is a

component of RNA polymerases I, II, and III and plays an important role in transcription [5], the main protein-coding transcript constituted ~50% in healthy *in vivo* liver but was completely replaced by other protein-coding and non-coding transcripts in all *in vitro* systems. Another example was *GOLGA8B* (Golgin subfamily A member 8B), for which non-coding transcripts in the cell models replaced the major protein-coding transcript (~ 60% in healthy *in vivo* liver), predominantly. Importantly, both these genes were not differentially expressed on the gene level, however, presented differential transcript usage. Numerous such cases were identified, thus highlighting the added benefits of performing a transcript-based analysis.

The PHH and 3D liver microtissues exhibited high similarity to the *in vivo* liver in most analyses performed. Each cell model demonstrated high similarity across certain pathways, e.g. HepG2 for DNA repair pathways. However, no cell model (among the cell models tested) could be presented as the "go-to" option for all studies. Through this study, an extensive data resource of differentially expressed genes, differential usage transcripts, pathway coverage for DEGs, non-DEGs, and non-DEGs^{DTU-} has been created that could be used to make a selection for processes, pathways, or genes of interest. In addition, the selection of the cell model for a study might also be influenced by various other factors such as cost, maintenance, and availability of expertise.

Different transcripts-same function

A gene can code for multiple transcripts due to the presence of various transcription start sites and alternative splicing [6, 7]. The analyses of the transcriptomics data have evolved predominantly centering the gene and only lately, the transcripts. While the gene-centered approaches mask any changes occurring on the transcript level, the transcript-based analysis overlooks the functional similarities of the transcripts. Functional similarity for the protein-coding transcripts being their ability to produce the same or similar proteins (and hence perform the same functions) and for the non-coding transcripts being their ability to bind the same targets and trigger the same/similar responses. Functionally similar transcripts might originate from the same or different genes and possess a similar coding sequence (CDS). From the similar CDS, courtesy of degeneracy of the codon, functionally similar proteins are translated.

When multiple transcripts are capable of coding for the same or similar proteins, an ensemble of expression of all such transcripts should be used to define the total amount of function, rather than analyzing individual transcripts or genes. In **Chapter 3**, this notion of different protein coding transcripts capable of coding for the same or similar proteins was explored and, to study its impact on RNA-Seq data, resulted in the development of FuSe (**Fu**nctional grouping of RNA-**Se**q data). FuSe is a command-line based tool that allows the user to search for protein-coding transcripts that would code for the same protein. For finding the functionally similar protein-coding transcripts, their amino acid sequence, secondary and super secondary structures were compared and a scoring scheme was developed to score the features. Two different scores were designed: a discovery score (DS, lenient) and a knowledge score (KS, stringent). Using these scores, flexibility is provided to search for transcripts with a certain amount of similarity and the features to be considered while calculating the similarity. It was also observed that the transcripts, which code for the same or similar proteins, originated from distinct genes as well (at KS≥95, ~78.5% from the same gene, ~15.5% from same gene family, ~2% from different gene family, and ~4% from genes with undefined family). For 60% of genes, the longest protein-coding transcript was the main form and the smaller protein-coding transcripts exhibited a different function, emphasizing that the major (known) function of a gene is not its only function.

FuSe was applied to liver cell models exposed to APAP (acetaminophen, therapeutic and toxic doses) and its equivalent controls (untreated and DMSO). It was seen that some of the important characteristics of APAP exposure were captured by FuSe while being completely missed by conventional RNA-Seq data analyses. One of the important examples was the *GBE1* expression. With conventional analyses, at the toxic APAP dose, *GBE1* was shown to be upregulated, implying that the glycogen accumulation in the liver has increased. However, glycogen depletion is considered as one of the early biomarkers of acetaminophen-induced hepatotoxicity [8] and through the application of FuSe, the expression of GBE1 was correctly quantified by accommodating other transcripts coding for the same protein.

Transcript biomarkers – a potent tool

Biomarkers are an important aspect of biomedical research. The different classes of biomarkers are used to address various aspects of a disease, for instance,

susceptibility, presence, progression, effects of treatment, or chances of reoccurrence. Predominantly, proteins are used as biomarkers, however, with the advancements in NGS technologies, the search for gene biomarkers has also accelerated [9-11]. The transcriptome landscape changes rapidly under exposure to drugs, alterations in the external environment, and lifestyle changes. The subsequent changes in the proteome are delayed due to the intermediate processes of translation and post-translational modifications of the protein. At the same time, as previously mentioned, the heterogeneity in gene expression makes them a poor choice for biomarkers. To address these challenges, we investigated transcript expression in **Chapter 4** to demonstrate if transcripts can project as better biomarkers. The RNA-Seq data for hepatocellular carcinoma (HCC) cell models was analyzed using machine learning to find transcript biomarkers. The HCC cell models selected for this study presented distinct genome-scale variations, for instance, HuH7.5.1 is homozygous for TP53 p.Tyr220Cys [12, 13], HKCI-4 has lost chromosome Y and aberrations in chromosome 10 [14], and PLC/PRF/5 contains at least 7 copies of integrated HBV genomes [15]. It was important to make a collection of such cell models to accommodate all possible variations in HCC.

From the analysis of the HCC cell models, three transcripts (two protein coding: PARP2-202 and SPON2-203 and one non-coding: CYREN-211) were selected through recursive feature elimination with an accuracy of 0.97 and kappa of 0.93. Consequently, it was also demonstrated that when the transcriptomics data were analyzed separately for protein-coding and non-coding transcripts, a comparatively lower accuracy was observed compared to when they were taken together (AUC-ROC, sensitivity, specificity, MCC, and informedness were all ~1 for the random forest, Naive Bayes, and Support vector machine). Additionally, the comparison of classical serum biomarkers (on the transcript level) was also performed and a comparatively lower accuracy was observed for them.

From the analyses of the whole transcriptomics data, protein-coding and non-coding transcripts were identified as HCC biomarkers. At the same time, comparatively lower accuracy for the split data (protein coding only, or non-coding only) establishes that both protein-coding and non-coding transcripts possess important biological information. Moreover, their expression changes under internal and external influences. This work is purely computational and no known relationship of the identified biomarkers (PARP2-202, SPON2-203, and CYREN-211) to the HCC could be found in the literature. Hence, before advancing further, these

results need to be validated in the lab. Regardless, the machine-learning pipeline developed in this study can be applied to any RNA-Seq data study aimed at finding transcript biomarkers.

Protein complexes – the multi-molecule machineries

Protein complexes are the multi-molecule machineries of the biological system. They are involved in several functions and are part of various cellular structures. Complex Portal, one of the manually curated protein complex database, currently holds information on 779 complexes. However, the knowledge and understanding of the protein complexes are limited. To enhance our understanding of these protein complexes, the protein complex expression and their assembly order were investigated in **Chapter 5,** using time-series RNA-Seq data from untreated liver spheroids *in vitro* cell models.

In this study, RNA-Seq data from eight-time points after initiating the cell culture for baseline liver microtissues was used. The subsequent time points represent the growth of the cell/tissue culture at baseline. For the calculation of the protein complex expression, the lowest expressed subunit was identified after adjusting the expression of the subunits for their respective stoichiometry. It was observed that at different time points, different subunits of the complex represented the lowest expressed subunit. For instance, SMAD2-SMAD3-SMAD4 complex (CPX-1) expression is defined by SMAD2 for time points 2, 8, and 240hours, SMAD3 for 168 hours, and SMAD4 for 0, 24, 72, and 336 hours. If the expression of the individual subunits was compared, up or downregulation was observed, however, the complex expression showed insignificant perturbation. A higher upregulation for SMAD3 was observed at time point 72 and 240 hours as compared to other time points. Similarly for SMAD4, 168 hours exhibited upregulation compared to all other time points. If only these subunits were compared, as for SMAD3 that is an intracellular signal transducer and transcriptional modulator [16], it would imply an increase in transcription. Similarly, the upregulation of SMAD4 implies an escalation in the TGF-mediated signaling [17]. However, the complex demonstrated insignificant perturbation in expression over time, thus implying that none of these functions were upregulated as shown by subunits. It is important to mention that the complex expression calculated from the RNA-Seq data is not an absolute

measure of the final assembled complex in the biological system, and it only provides a metric to compare the functional changes.

Furthermore, the protein complexes are formed in a stepwise manner, binding one or a few subunits at each step. A stepwise binding allows the forming complex to change in confirmation and to expose hidden binding regions for other incoming subunits. By means of the application of the dynamic Bayesian networks (DBN) on the RNA-Seq time-series data, the order of assembly of the protein complexes was predicted by analyzing the interaction over consecutive time points. An important feature of the DBN is that the nodes only have children in further time points i.e. a node from time point 't' will only have a child in time point t+1, t+2, … t+n but not in t-1, t-2, … t-m (t: time point, n, m: positive integers). The subunits that exhibit interaction at the initial time points would come together earlier as compared to subunits that show interaction at later time points. The order of assembly for Laminin211-nidogen complex (CPX-1282) was predicted from analyzing the edges (interactions) between the nodes (subunits) over consecutive time points.

The elucidation of the order of assembly may create opportunities to develop novel drug targets. The binding of the subunits creates and opens up hidden binding sites due to conformational changes in the formation of the protein complex [18, 19]. The conformational changes occur to achieve the minimum energy state. These conformational changes can thus present new binding sites for the drugs.

Challenges

The major challenge in the transcript-based analyses of the data is the unavailability of transcript-function databases. Currently, biological databases (for GO, pathways, function, interaction, etc.) predominantly store and present data in terms of genes and/or proteins. The transcripts are un-mappable on several databases or are mapped with significant loss of data. We addressed this issue by identifying the changes at the transcript level (DTU) and then refining the list of non-DEGs by removing the genes that have a transcript exhibiting differential usage. Through this approach, we were able to capture the changes at the transcript level and study the function at the gene level. However, the non-existence of an exhaustive transcript-function database limited the outcome of the study. Additionally, we encountered multiple challenges making the editor and the reviewers agree on the transcript-based analyses presented in Chapter 2. However, various RNA-Seq

studies have illustrated the significance of transcript-oriented analyses; the major proportion of the RNA-Seq community still prefers gene-expression based analyses.

Furthermore, while working to find the transcripts that code for the same or similar proteins, the selection of the features of the proteins to be used was a major challenge. The complete and purified 3D structure of many human proteins is not available [20] and hence could not be used to compare the proteins. We resorted to amino acid sequence and secondary structures of the proteins for establishing similarity between the proteins. However, due to the lack of any pre-existing approach to assemble results from sequence and secondary structure similarities, novel approaches were developed.

Finally, for the work on the protein complexes, we realized that only a handful of protein complexes are known (Complex Portal: 779 [21]) and out of them, one-third of the complexes do not have defined stoichiometry. The stoichiometry is important to know for the calculation of the complex expression, so it was assumed that the stoichiometry of such complexes was one for all subunits. This work would improve our understanding of the biological interpretation of the RNA-Seq data; however, the lack of information on the different protein complexes limits its applicability.

Limitations

In Chapter 2, the comparison is based on baseline expression of the cell models and hence it cannot illustrate the response when exposed to certain drugs, culture conditions, and other controllable or uncontrollable parameters. However, the study presents the response of the cell models concerning certain processes and can be used as the base for future research and interpretations. Further in Chapter 3, while FuSe identifies transcripts coding for the same or similar proteins, it does not cover the non-coding transcripts. The involvement of the non-coding transcripts in various regulation processes is well established [22-25] and hence would need an analogous similarity finding tool or utility. Next in Chapter 4, the identification of the transcript biomarkers for HCC using various machine-learning algorithms is illustrated. However, the predicted biomarkers illustrate high accuracy in their ability to distinguish the healthy and HCC cell models through computational approaches, their validation in the HCC patients is yet to be done. A study thoroughly investigating the predicted biomarkers in HCC patients is required

to establish confidence in the *in silico* findings. Lastly, the work on the prediction of the protein complex expression and its assembly order could not be verified due to the unavailability of wet-lab findings.

Future work

In this thesis, across the various research chapters, multiple ideas have been explored and developed, for instance finding the transcripts coding for the same or similar proteins (Chapter 3), identification of transcript-based biomarkers for HCC as an application of a novel ML-based approach (Chapter 4), and prediction of protein complex order of assembly using DBN (Chapter 5). For each of these research objectives, there is immense scope for further development of the idea and investigating the outcomes.

As in Chapter 3, the identification of the similar protein coding transcripts was done using the amino acid sequence, secondary and super-secondary structures. However, various other attributes of the proteins could also be added to this comparison to make it comprehensive. Some of the features of the proteins that should be included in the comparison are charge of the proteins, stability, degradation, and half-life. The protein charge can be calculated from the assessment of the charged amino acids in the protein and will play a vital role in the function of the protein by allowing or preventing certain interactions [26, 27]. Protein stability is another important feature that can contribute to assessing the similarity between the proteins. Through various computational approaches, protein stability can be predicted [28-30]. The degradation and half-life of the transcripts and proteins can further add important information on their life span in the cell and hence could be used to predict the expression at a given time [31, 32]. The functionality comparison could also be developed for the non-coding transcripts to assess the similarity of function across various non-coding transcripts. The similarity in function of the non-coding transcripts could be achieved by analyzing their share binding target motifs.

Furthermore, the identification of the biomarkers from the transcriptomics data using ML for HCC is an in-silico study (Chapter 4) that needs a follow-up validation study. A population study to sequence healthy and HCC patients' liver samples needs to be done. Targeted RNA-Seq such as TempO-Seq or Targeted Enrichment RNA-Seq should be performed to target the specific transcripts that are identified

as HCC biomarkers. Such a study would establish if the predicted transcript biomarkers for HCC are efficient in distinguishing the healthy and diseased states in the real-world scenario as well.

Lastly, the work on the prediction of the protein complex's assembly order as discussed in Chapter 5 can be developed further, first by refining the selection of the protein-complexes based on the information of stoichiometry, identifier mapping, and expression of the transcripts. Additionally, the use of FuSe, the tool developed in Chapter 3, to calculate the expression of the transcripts coding for the same or similar proteins, could further enhance the prediction of the protein complex's assembly order. Additionally, the current generation of the DBNs is based on connections between individual subunits over consecutive time points and these should be changed to the nascent protein complex that is defined by the subunits bound at each time point as predicted by DBN. It would mean that the DBNs should be computed iteratively by eliminating the first time point in each step and replacing the bound subunits with the nascent protein complex. These in-silico predictions should be made the basis for the electrospray mass spectrometry, co-purification, co-crystallization, Yeast2Hybrid, and genetic interactions experiments to study the binding of proteins and hence the order of assembly of the complexes. The new knowledge gained from these studies can be used in improving the in-silico predictions.

Conclusions

Novel analyses and applications of the transcript expression data from RNA-Seq have been illustrated in this thesis. These studies provide an opportunity to further explore and understand biological complexities. Through the transcript-based analyses, the heterogeneity in the gene expression as obtained from RNA-Seq data and the ability of different transcripts to code for the same proteins are addressed. A transcript-based biomarker identification approach and a method to predict the protein-complex assembly order are presented. This work on transcript-based analyses of the RNA-Seq data provides another dimension to the data analyses and aims to better understand the biological processes.

References

1. DelRaso NJ: **In vitro methodologies for enhanced toxicity testing**. *Toxicol Lett* 1993, **68**(1-2):91-99.
2. Godoy P, Hewitt NJ, Albrecht U, Andersen ME, Ansari N, Bhattacharya S, Bode JG, Bolleyn J, Borner C, Boettger J: **Recent advances in 2D and 3D in vitro systems using primary hepatocytes, alternative hepatocyte sources and non-parenchymal liver cells and their use in investigating mechanisms of hepatotoxicity, cell signaling and ADME**. *Archives of toxicology* 2013, **87**(8):1315-1530.
3. Radley AH, Schwab RM, Tan Y, Kim J, Lo EKW, Cahan P: **Assessment of engineered cells using CellNet and RNA-seq**. *Nature Protocols* 2017, **12**:1089.
4. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: **The KEGG resource for deciphering the genome**. *Nucleic Acids Res* 2004, **32**(suppl_1):D277-D280.
5. Kershnar E, Wu SY, Chiang CM: **Immunoaffinity purification and functional characterization of human transcription factor IIH and RNA polymerase II from clonal cell lines that conditionally express epitope-tagged subunits of the multiprotein complexes**. *J Biol Chem* 1998, **273**(51):34444-34453.
6. Roth MJ, Forbes AJ, Boyne MT, 2nd, Kim YB, Robinson DE, Kelleher NL: **Precise and parallel characterization of coding polymorphisms, alternative splicing, and modifications in human proteins by mass spectrometry**. *Mol Cell Proteomics* 2005, **4**(7):1002-1008.
7. Karlsson C, Malmström L, Aebersold R, Malmström J: **Proteome-wide selected reaction monitoring assays for the human pathogen Streptococcus pyogenes**. *Nat Commun* 2012, **3**(1301).
8. Gautam R, Chandrasekar B, Deobagkar-Lele M, Rakshit S, Kumar B. N V, Umapathy S, Nandi D: **Identification of Early Biomarkers during Acetaminophen-Induced Hepatotoxicity by Fourier Transform Infrared Microspectroscopy**. *PloS one* 2012, **7**(9):e45521.
9. Akond Z, Alam M, Mollah MNH: **Biomarker Identification from RNA-Seq Data using a Robust Statistical Approach**. *Bioinformation* 2018, **14**(4):153-163.
10. Han J, Chen M, Wang Y, Gong B, Zhuang T, Liang L, Qiao H: **Identification of Biomarkers Based on Differentially Expressed Genes in Papillary Thyroid Carcinoma**. *Sci Rep-Uk* 2018, **8**(1):9912.
11. Shigemizu D, Mori T, Akiyama S, Higaki S, Watanabe H, Sakurai T, Niida S, Ozaki K: **Identification of potential blood biomarkers for early diagnosis of Alzheimer's disease through RNA sequencing analysis**. *Alzheimer's Research & Therapy* 2020, **12**(1):87.
12. Fostira F, Konstantopoulou I, Mavroudis D, Tryfonopoulos D, Yannoukakos D, Voutsinas GE: **Genetic evaluation based on family history and Her2 status correctly identifies TP53 mutations in very early onset breast cancer cases**. *Clin Genet* 2015, **87**(4):383-387.
13. Pittman AM, Lage MD, Poltoratsky V, Vrana JD, Paiardini A, Roncador A, Cellini B, Hughes RM, Tucker CL: **Rapid profiling of disease alleles using a tunable reporter of protein misfolding**. *Genetics* 2012, **192**(3):831-842.
14. Pang E, Wong N, Lai PBS, To KF, Lau WY, Johnson PJ: **Consistent chromosome 10 rearrangements in four newly established human hepatocellular carcinoma cell lines**. *Genes, Chromosomes and Cancer* 2002, **33**(2):150-159.
15. Twist EM, Clark HF, Aden DP, Knowles BB, Plotkin SA: **Integration pattern of hepatitis B virus DNA sequences in human hepatoma cell lines**. *Journal of Virology* 1981, **37**(1):239-243.
16. Zhao Y, Shi X, Ding C, Feng D, Li Y, Hu Y, Wang L, Gao D, Tian X, Yao J: **Carnosic acid prevents COL1A2 transcription through the reduction of Smad3 acetylation via the AMPKα1/SIRT1 pathway**. *Toxicol Appl Pharmacol* 2018, **339**:172-180.
17. Yang Y, Cui J, Xue F, Zhang C, Mei Z, Wang Y, Bi M, Shan D, Meredith A, Li H *et al*: **Pokemon (FBI-1) interacts with Smad4 to repress TGF-β-induced transcriptional responses**. *Biochim Biophys Acta* 2015, **3**:270-281.
18. Novak P, Havlicek V, Derrick PJ, Beran KA, Bashir S, Giannakopulos AE: **Monitoring conformational changes in protein complexes using chemical cross-linking and Fourier transform ion cyclotron resonance mass spectrometry: the effect of calcium binding on the calmodulin-melittin complex**. *Eur J Mass Spectrom* 2007, **13**(4):281-290.
19. Orellana L: **Large-Scale Conformational Changes and Protein Function: Breaking the in silico Barrier**. *Frontiers in Molecular Biosciences* 2019, **6**(117).
20. Liu Z, Li Y, Han L, Li J, Liu J, Zhao Z, Nie W, Liu Y, Wang R: **PDB-wide collection of binding data: current status of the PDBbind database**. *Bioinformatics* 2015, **31**(3):405-412.
21. Meldal BHM, Forner-Martinez O, Costanzo MC, Dana J, Demeter J, Dumousseau M, Dwight SS, Gaulton A, Licata L, Melidoni AN *et al*: **The complex portal - an encyclopaedia of macromolecular complexes**. *Nucleic Acids Res* 2014, **43**(D1):D479-D484.

22. Briata P, Gherzi R: **Long Non-Coding RNA-Ribonucleoprotein Networks in the Post-Transcriptional Control of Gene Expression**. *Noncoding RNA* 2020, **6**(3).
23. Arabian M, Mirzadeh Azad F, Maleki M, Malakootian M: **Insights into role of microRNAs in cardiac development, cardiac diseases, and developing novel therapies**. *Iran J Basic Med Sci* 2020, **23**(8):961-969.
24. Guo X, Tan W, Wang C: **The emerging roles of exosomal circRNAs in diseases**. *Clin Transl Oncol* 2020, **15**(10):020-02485.
25. Saw PE, Xu X, Chen J, Song EW: **Non-coding RNAs: the new central dogma of cancer biology**. *Sci China Life Sci* 2020, **11**(10):020-1700.
26. Lošdorfer Božič A, Podgornik R: **pH Dependence of Charge Multipole Moments in Proteins**. *Biophys J* 2017, **113**(7):1454-1465.
27. Højgaard C, Kofoed C, Espersen R, Johansson KE, Villa M, Willemoës M, Lindorff-Larsen K, Teilum K, Winther JR: **A soluble, folded protein without charged amino acid residues**. *Biochemistry* 2016, **55**(28):3949-3956.
28. Montanucci L, Capriotti E, Frank Y, Ben-Tal N, Fariselli P: **DDGun: an untrained method for the prediction of protein stability changes upon single and multiple point variations**. *BMC Bioinformatics* 2019, **20**(14):335.
29. Yang Y, Ding X, Zhu G, Niroula A, Lv Q, Vihinen M: **ProTstab – predictor for cellular protein stability**. *BMC genomics* 2019, **20**(1):804.
30. Chen C-W, Lin M-H, Liao C-C, Chang H-P, Chu Y-W: **iStable 2.0: Predicting protein thermal stability changes by integrating various characteristic modules**. *Computational and Structural Biotechnology Journal* 2020, **18**:622-630.
31. Zhou P: **Determining protein half-lives**. In: *Signal Transduction Protocols.* Springer; 2004: 67-77.
32. Rahman M, Sadygov RG: **Predicting the protein half-life in tissue from its cellular properties**. *PloS one* 2017, **12**(7):e0180428-e0180428.

Impact

The primary goal of the research is to add to the existing knowledge and deepen our understanding of a topic. The primary analysis of the RNA-Seq data is gene-based, however, in this thesis, we showed that a transcript-based analysis approach produces better biological interpretations. At the same time, it allows for novel analysis of the RNA-Seq data generating new hypotheses and results.

The study of comparison of different liver cell models through RNA-Seq data resulted in the generation of an exhaustive resource outlining the similarities and differences of the cell models to liver biopsies, as presented in chapter 2 (1). The cell models *ex-vivo* lose various *in-vivo* characteristics (2, 3). However, the unavailability of a comprehensive comparison restricted the assessment of the changes. A thorough study of the liver cell models through RNA-Seq data at the gene and transcript level illustrated the need for moving to a transcript-based approach. The biological changes in the cell models were highlighted with more precision using the transcript expression. The data generated through this comparison will be beneficial in selecting the cell models for specific research questions – what pathways, processes, and/or traits are of interest? Using the comparison data available, informed decisions can be made. This work can be used as a template to compare cell models from other tissues and cell types as well. A resource of cell models defining how similar or different they are to the *in vivo* systems would be helpful to perform better research.

The knowledge of different transcripts originating from the same or different genes making the same protein always existed (4, 5). However, while analyzing the RNA-Seq data the focus was always on evaluating the expression of individual genes or transcripts. The assimilation of this concept of the same protein from different transcripts to the RNA-Seq data analysis resulted in the creation of FuSe (6). Through this approach, the RNA-Seq data analysis provided more information on the dynamics of the biological system. However, various layers of regulation are involved in making a protein from the mRNA, the grouped expression calculation gives the preliminary protein expression estimates. This can be used as a starting point to accommodate other regulations and reach to the protein expression values. The work emphasized that the biological systems try to achieve homeostasis by producing different transcripts that code for the same proteins. The changes in the type and expression of these transcripts can be attributed to the internal or external environment. The identification of such transcripts that are involved in coding for the same proteins calls for studying their evolutionary relationships.

Using machine learning (ML) approaches with the transcript expression data novel potent transcript biomarkers were identified (7, 8). The study focused on various HCC cell models and presented a group of protein coding and non-coding transcripts as the potent biomarkers for detection of HCC. The inclusion of non-coding transcripts in the biomarker discovery results emphasized their underlying role in disease progression. The added advantage of using transcripts is that they are produced before the proteins in the system and can help in the early detection of the diseases. The identification of the transcripts as biomarkers can have a major impact on future of biomarker research. Novel and more potent biomarkers can then be identified for diseases where early diagnosis is a major hurdle.

Lastly, evaluating the expression of the protein complexes from the RNA-Seq data provided more functional assessment of the biological system. From an entity-based analyses (gene or transcript), we could then define the amount of work achievable in the biological system. Additionally, the information of the assembly of the protein complexes elevated our understanding of the multi-molecule machinery and opened new doors for investigating novel drug targets. The application of dynamic Bayesian networks to the temporal RNA-Seq data helped in generating hypothesis for assembly of protein complexes. Though preliminary, this work has the capacity to control and eliminate various diseases occurring due to mis-assembly of the protein complexes. Moreover, information on the assembly would highlight the evolutionary information. What genes moved apart and what came closer over years of evolution (9).

We demonstrated that the study of RNA-Seq data is no longer limited to evaluating genes or transcripts. It can be used to assess the amount of protein that can be formed in the biological system, present transcripts as potent disease biomarkers, can help elucidate protein complex assembly and more.

References

1. Gupta R, Schrooders Y, Hauser D, van Herwijnen M, Albrecht W, ter Braak B, et al. Comparing in vitro human liver models to in vivo human liver using RNA-Seq. Archives of Toxicology. 2021;95(2):573-89.
2. Sachinidis A, Albrecht W, Nell P, Cherianidou A, Hewitt NJ, Edlund K, et al. Road map for development of stem cell-based alternative test methods. Trends in molecular medicine. 2019.
3. Albrecht W, Kappenberg F, Brecklinghaus T, Stoeber R, Marchan R, Zhang M, et al. Prediction of human drug-induced liver injury (DILI) in relation to oral doses and blood concentrations. Archives of toxicology. 2019;93(6):1609-37.

4. Mariño-Ramírez L, Kann MG, Shoemaker BA, Landsman D. Histone structure and nucleosome stability. Expert review of proteomics. 2005;2(5):719-29.
5. Hegde A. Ubiquitin-proteasome system and plasticity. 2009.
6. Gupta R, Schrooders Y, Verheijen M, Roth A, Kleinjans J, Caiment F. FuSe: A tool to move RNA-Seq analyses from chromosomal/gene loci to functional grouping of mRNA transcripts. Bioinformatics. 2020;1:7.
7. Johnson NT, Dhroso A, Hughes KJ, Korkin D. Biological classification with RNA-seq data: Can alternatively spliced transcript expression enhance machine learning classifiers? RNA (New York, NY). 2018;24(9):1119-32.
8. Akter S, editor A Data Mining Approach for Biomarker Discovery Using Transcriptomics in Endometriosis. 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2018: IEEE.
9. Marsh JA, Hernández H, Hall Z, Ahnert SE, Perica T, Robinson CV, et al. Protein complexes are under evolutionary selection to assemble via ordered pathways. Cell. 2013;153(2):461-70.

Acknowledgements

PhDs are tough, they are meant to be but with the right people around you, the ride becomes beautiful. I was lucky to be surrounded by wonderful people who made it seem like a gust of cold breeze on a hot summer day. I would like to extend my gratitude to multiple people who have in one way or the other supported this work or me through this journey.

First and foremost, I would like to thank **Prof. dr. Jos Kleinjans**, who trusted in me and offered me the Ph.D. position at the TGX. Your supervision helped in giving direction to my all over the place ideas. Your appreciation gave me the confidence to go further and to push my limits.

I fall short of words to thank **Dr. Florian Caiment** for his guidance and support. You are truly an amazing person and mentor. You always took the effort to explain the simplest to the toughest of the concepts, encouraged me to do better, and was always there when I fell. You gave me the freedom to explore my ideas and helped to develop analytical thinking however at the same time also showed me the way when I was lost. I loved the work meetings with you where we talked about work and also engaged in various other topics. I could not have asked for a better supervisor.

A heartfelt thanks to **Prof. dr. Theo**, **Dr. Danyel Jennen**, **Dr. Jacco Briedé**, **Dr. Simone Breda**, **Dr. Twan Beucken**, and **Dr. Jumamyrat Bayjanov** for constructive feedback on my work; it helped in shaping it into a better version. A special thanks to all the collaborators from **EU-ToxRisk**. The talks and discussions at the assembly meetings, conferences, and workshops were hugely beneficial.

Tim is such a wonderful soul; once I said I feel like having butterfly cookies and he baked them 😊. He took time to debug my code and pointed out beginner's mistakes but never complained. Thank you for arranging the Ph.D. quiz via Zoom in this lockdown. **Marcha** is my go-to person for almost everything. Where to find this data? How to write that report? What to use for my thesis cover? And she has always answered me with a big smile on her face. **Terezinha** always shared her knowledge and answered my doubts. Your delicious prawn recipe has stayed with me. **Evelyn** introduced me to the *Dutch directness*. Though it felt strange in the beginning, she prepared me for the years to come. **Julian** has been a great person. He helped me move my house and carried all the boxes and bags. The wonderful chessboard that you gave me is a much-cherished possession. **Héloïse's** laughter from the next door made sure I was never asleep in the office and was always

working on my deadlines. In my initial days when I was lost with the official dutch letters, **Jarno** always extended a helping hand. He introduced me to the luxurious dutch lunch, what a treat that was! **Jian's** friendly nature is very comforting. Every now and then her messages asking for work and life feels so good. Also, thank you for sending that wonderful gift to Munich. **Daniella** – my Dutch learning partner; I still do not know how I passed A1. With the arrival of **Juan** to the room, it quickly changed from the always quiet to not-so-quiet any more room. All the chats and discussions were so fun-filled. The ever amazing **Nhan** was so good with her occasional witty one-liners. The batman coffee mug and self drawn card from **Marta** showed how much she cared. **Julia** kept spreading her positivity. **Manon** was the powerhouse of the room; handling lab animals, coding, spinning trainer, and whatnot. If I could be half as energetic as you… . **Maria**, **Almudena** thank you all for the wonderful conversations. I loved the cricket-analysis with **Yaseer**, maybe we can start our post-match shows. All the work, lunches, chats, planks, wall sits, and parties were amazing because of all you guys. Thank you for making my stay memorable.

The ride would not have been so smooth if the cumbersome administrative works weren't taken care of by **Christa**, **Rob**, and **Rene**. Much obliged! One of the most wonderful gesture was when Rene came to pick me up at the Maastricht station on my arrival. I do not know how I would have managed to find my way home that day with a switched off mobile. Christa has been no less than a guardian angel making sure everything is perfectly arranged. She even gave me her mobile number when I was traveling for the first time to a conference, so that if there were any problems, I could reach her even after work. And Rob was always there making sure all the bills and costs were paid back so we do not have to worry about anything.

To **Yannick**, **Duncan**, and **Marcel**, thank you for taking care of the lab work and producing high-quality data. Also, Marcel's efforts in making sure that the sitting position, table, and chair were all good, made sure my neck and back have got better. The closed-door conversations with Yannick, Tim, and Juan; okay let's not talk about them here. The most important thing for a bioinformatician is the servers, thanks for keeping them up and running: **Tom**, **Kevin**, and **Sean**.

The little Indian community that I have found in Maastricht never made me feel that I was away from home. **Shanmuk** and **Sarvani**: the home where there is no dearth of love and food; **Krishna** and **Ritu**: food, fun, and games; **Harpreet** and **Nitti**: the Wikipedia of knowledge with overflowing generosity; **Pankaj** and **Vini**: one-

stop-shop for relaxing evenings and Jagjit Singh; **Xan**: one woman army; **Jyothi** and **Pierre** (the desi angrez): the ever welcoming. Thank you for being a part of my journey and spreading nothing but love. **Nitish**, my friend my brother, you have always been a breeze of cool air. Thanks for your support all through.

A special mention goes to **Dr. Dipankar Sengupta**, if not for him I would not have pursued my dream to get a Ph.D. His continuous encouragement made sure I reach where I am today. Mere words cannot express how much gratitude I feel. A good landlord and landlady are not a myth; I feel blessed to meet **Raijmond** and **Liliane**. All the delightful dinners and chats defined the stay at Demertstraat.

Abhijeet is the person who made me believe that I can code and has always been there when I got stuck in technical know-how. You are the friend (and Stack overflow) I always needed. My childhood friends **Arjun**, **Mayur**, and **Akshay** kept feeding me my fair dose of laughter through calls and messages. Whenever I felt lost, Arjun made sure he has a speech ready for me to lift me. The long list of uncles, aunts, and cousins made sure I paid equal attention to my health as to my work.

Badimumma-daddy and **Mumma-papa**, who have loved me unconditionally and have always showered their blessings on me. I can fly high because I know you are holding the string and you will never let me fall. Your sacrifices and hard work made my life a walk in the garden. Thank you for everything 🙏. My younger brother **Pinku** took care of all the ups and downs of the family and made sure that the tides do not reach me. *He is the Hero every family needs*. May God gives you all the happiness! My **mother-in-law** and **father-in-law** are one of the most delightful people and always manage to cheer me up. And finally, my beloved wife **Shaveta**, who sacrificed a lot for my Ph.D. and never complained (okay, sometimes 😊). Thank you for standing by my side and bearing with my frustration and anger, and pulling me out of the darkness, every time I was down. You have been the best audience to my work presentations and the best partner I could have asked for. An eternity will be too little to spend with you.

The four strongest pillars that kept me standing all these years: my promotion team, my colleagues, my friends, and my family.

Thank you, everyone!

Curriculum vitae

Rajinder Gupta was born on 4th April 1989 in Jammu, India. He started his bachelor's degree in 2007 in Biotechnology from DYPBBI University Pune, India. For the bachelor's thesis, he worked on the toxicity effects of novel plant extracts. He then joined JUIT, Solan to pursue a master's degree in Computational Biology. His master's thesis involved the exploration of the networks in cancer pathways. He obtained the gold medal for scoring merit in the class. In 2014, he joined National Agri-food Biotechnology Institute, Mohali as a Junior research fellow. There he worked on the development of a Universal biomolecular relationship database, extracting and collecting biological interaction data from databases and scientific publications. In May 2016 he joined the department of Toxicogenomics at Maastricht University as a Ph.D. student. He was hired under the EU-ToxRisk project where he was given the opportunity to analyze RNA-Seq data from various human cell models through established approaches and to explore new avenues. Under the guidance of Prof. dr. Jos Kleinjans and Dr. Florian Caiment, the work resulted in the thesis titled "Beyond gene expression: Novel methods and applications of transcript expression analyses in RNA-Seq". Currently, he works as a PostDoc in Emmy Noether Group for Computational Microbiome Research under Dr. Melanie Schirmer. He will be investigating and developing ways to apply novel RNA-Seq analysis approaches developed for human data to microbes.

List of Publications

Published

- **Gupta, Rajinder**, Yannick Schrooders, Duncan Hauser, Marcel van Herwijnen, Wiebke Albrecht, Bas Ter Braak, Tim Brecklinghaus et al. "*Comparing in vitro human liver models to in vivo human liver using RNA-Seq.*" Archives of Toxicology (2020): 1-17.
- **Gupta, Rajinder**, Yannick Schrooders, Marcha Verheijen, Adrian Roth, Jos Kleinjans, and Florian Caiment. "*FuSe: A tool to move RNA-Seq analyses from chromosomal/gene loci to functional grouping of mRNA transcripts.*" Bioinformatics (2020).
- Attoff, Kristina, Ylva Johansson, Andrea Cediel Ulloa, Jessica Lundqvist, **Rajinder Gupta**, Florian Caiment, Anda Gliga, Anna Forsby *"Acrylamide alters CREB and retinoic acid signaling pathways during differentiation of the human neuroblastoma SH-SY5Y cell line."* Sci Rep 10, 16714 (2020). https://doi.org/10.1038/s41598-020-73698-6
- **Gupta, Rajinder**, and Shrikant S. Mantri. "*Biomolecular Relationships Discovered from Biological Labyrinth and Lost in Ocean of Literature: Community Efforts Can Rescue Until Automated Artificial Intelligence Takes Over.*" Frontiers in Genetics 7 (2016): 46.
- Gurjar, Anoop Kishor Singh, Abhijeet Singh Panwar, **Rajinder Gupta**, and Shrikant S. Mantri. "*PmiRExAt: plant miRNA expression atlas database and web applications.*" *Database* 2016 (2016).
- Sehgal, Manika, **Rajinder Gupta**, Ahmed Moussa, and Tiratha Raj Singh. "*An integrative approach for mapping differentially expressed genes and network components using novel parameters to elucidate key regulatory genes in colorectal cancer.*" Plos one 10, no. 7 (2015): e0133901.

In process

- **Gupta, Rajinder**, Jos Kleinjans, Florian Caiment *"Identifying the transcriptomics biomarkers for hepatocellular carcinoma (HCC) using machine learning" Minor Revision,* BMC Cancer
- **Gupta, Rajinder**, Jos Kleinjans, Florian Caiment *"Unravelling the protein complex assembly using transcriptomics and dynamic Bayesian networks" Submitted to* Proteins: Structure, Function, and Bioinformatics
- Vidya Chandrasekaran, **Rajinder Gupta**, Elisabeth Feifel, Georg Kern, Judith Lechner, Florian Caiment, Jos CS Kleinjans, Gerhard Gstraunthaler, Paul Jennings and Anja Wilmes *"Generation and characterisation of iPSC-derived renal proximal tubule-like cells with extended stability" Submitted to Scientific Reports*

List of talks and abstracts

Invited talks

July 2020 — FuSe: A tool to move RNA-Seq analyses from chromosomal/gene loci to functional grouping of mRNA transcripts
Event: International Webinar on Biotechnology, Bioinformatics And Natural Products in Health Care at the Department of Biotechnology & Bioinformatics, Sambalpur University, Jyoti Vihar, Odisha, India

Aug, 2018 — Comparing *in vitro* human liver models to *in vivo* human liver using RNA-Seq
Event: EUROTOX 2018

Conference talks

Feb 2019 — Comparing *in vitro* human liver models to *in vivo* human liver using RNA-Seq
Event: EU-ToxRisk, General Assembly

Conference posters/abstracts

Nov, 2019 — FuSe: Moving RNA-Seq analyses from chromosomal/gene loci to functional grouping of mRNA transcripts
Event: byteMAL, 2019

Mar, 2019 — Improving the interpretation of omics analyses
Event: Revolutionizing Next-Generation Sequencing, Antwerp

Feb, 2019 — Comparing different *in vitro* liver models using RNA-Seq
Event: EU-ToxRisk, General Assembly

Nov, 2018 — Improving the interpretation of omics analyses
Event: GROW Science Day, UM

Feb, 2018 — Sequencing EU-ToxRisk human *in vitro* cell models for comparing expression of signaling pathways at baseline
Event: EU-ToxRisk, General Assembly

Nov, 2017 — The importance of isozymes in pathway analyses
Event: GROW Science Day, UM

Mar, 2016 — Connections: Universal Bio-molecular relationship database
Event: National Symposium on Computational Systems Biology, JUIT, Solan

Feb, 2014 — Deciphering Biological networks: Clues for cure
Event: Biorhythm under the aegis of Department of Biotechnology India

Oct, 2013 — Identification and analysis of network motifs in human disease specific pathways applying top down Systems Biology approach
Event: Virtual Conference, Bioinformatics to Systems Biology

Sep, 2013 — Annotation of Biological Networks through Network Motifs using Top Down approach
Event: 3rd IFIP International Conference of Bioinformatics, MANIT, Bhopal

Abbreviations

APAP	Acetaminophen or Paracetamol
AUC-ROC	Area under the curve-receiver operating characteristics
DEGs	Differentially expressed genes
DTU	Differential transcript usage
ENA	European Nucleotide Archive
FN	False negative
FP	False positive
FuSe	Functional grouping of transcripts for RNA-Seq analyses
HCC	Hepatocellular Carcinoma
hPCLiS	Human precision-cut liver slices from HCC patients
iPSC	induced pluripotent stem cells
KNN	K-Nearest neighbors
MCC	Mathew's correlation coefficient
ML	Machine learning
NB	Naïve Bayes
NGS	Next-generation sequencing
NNET	Neural Networks
Non-DEGs	Non-Differentially expressed genes
Non-DEGs^{DTU-}	Non-Differentially expressed genes without transcripts having differential usage
PHH	Primary human hepatocytes
RF	Random Forest
RNA-Seq	RNA sequencing
SFPGs	Similar function protein groups
SVM	Support vector machine
TN	True negative
TP	True positive